1978

Constructive optimization for infinite-dimensional problems in probability and statistics

Geung-Ho Kim
Iowa State University

Constructive optimization for infinite-dimensional problems in probability and statistics

by

Geung-Ho Kim

A Dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of

The Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Co-majors: Statistics
Industrial Engineering

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Departments

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1978

TABLE OF CONTENTS

I. INTRODUCTION

This work is about a certain "constructive" programming approach and its applications to some infinite-dimensional constrained optimization problems arising in probability and statistics. This constructive approach exploits the algorithmic nature inherent in the notion of weak duality of mathematical programming, and, when necessary, has recourse to the elementary but very useful notion of subgradient or support from convex analysis. In this approach, optimization proceeds by verifying a trial solution. As for the issue of existence of such solutions is concerned, we maintain the view that nonuniform treatment of the issue may expedite the problem-solving procedure. In fact, in all applications below the handling of the issue of existence is tailored to the specific nature of the individual application problem; explicitly settled independent of the programming argument for some problems or implicitly settled at the stage of problem formulation for others.

In view of this pragmatic orientation, our treatment of abstract programming duality, minimizing structural assumptions typical of those

treatments--for example in Rockafellar [51], Van Slyke and Wets [59], and Varaiya [60]--that are oriented toward general conditions for existence of optimal solution including "strong" duality, invokes no topological space setting, in the spirit, for example, of Duffin [17].

The verifying of a trial solution is based on the notion of weak duality: In order to verify that x^* solves the

Problem P: $\min h(x)$, $x \in R$,

one finds a

Problem D: $\max k(y)$, $y \in S$,

such that

$$h(x) \geq k(y), \quad x \in R, \quad y \in S, \tag{1.1}$$

i.e., a problem D weakly dual to the problem P; then find a y^* such that

$$h(x^*) = k(y^*). \tag{1.2}$$

So doing shows that

$$h(x^*) = \min_{x \in R} h(x) = \max_{y \in S} k(y) = k(y^*); \tag{1.3}$$

i.e., shows that x^* and y^*, respectively, solve P and D, and that P and D are in fact (strongly) dual. Note that the extremizing directions in

P and D are interchangeable with the reverse of the inequality in 1.1, as in the following example, drawn from Chapter V.

Let $F(\cdot)$ and $G(\cdot)$, respectively, denote the partially known distribution of the strength of an object and the known distribution of the stress to which the object is subject. Corresponding to the problem of finding the maximum probability of reliable performance with the first two moments for $F(\cdot)$ specified, namely,

$$\text{Problem } P_1: \quad \max h_1(F), \ F\epsilon\Phi,$$

where $h_1(F) \equiv \int G(t)dF(t)$, and $\Phi \equiv \{F \mid \int t^i dF(t) = b_i, \ i = 1,2\}$, we find a

$$\text{Problem } D_1: \quad \min k_1(\lambda), \ \lambda \equiv (\lambda_0,\lambda_1,\lambda_2)\epsilon\Lambda,$$

where $k_1(\lambda) \equiv \lambda_0 + \lambda_1 b_1 + \lambda_2 b_2$, and $\Lambda \equiv \{\lambda\epsilon E_3 \mid \lambda_0 + \lambda_1 t + \lambda_2 t^2 \geq G(t), \ \forall t\}$, such that

$$h_1(F) \leq k_1(\lambda), \ F\epsilon\Phi \text{ and } \lambda\epsilon\Lambda. \tag{1.4}$$

Based on 1.4, the verifying of the equality

$$\int[\lambda_0^* + \lambda_1^* t + \lambda_2^* t^2 - G(t)]dF^*(t) = 0, \tag{1.5}$$

analogous to 1.2, with a candidate pair (F^*,λ^*) amounts to constructing an optimal pair (F^*,λ^*) satisfying 1.5 ab initio.

Relation 1.5 leads to the construction of a two-point mass distribution F^*. As indicated in Chapter V, in general the optimal solution for this type of linear problem is a discrete mass distribution. It seems remarkable that, in contrast to this discrete optimal, the optimal solutions for the convex problems to be treated later tend to spread their mass; i.e., tend to form a continuous distribution. This fact will be exemplified later on.

A variation in emphasis occurs when both P and D are given and trial optima x* and y* are simultaneously to be verified. A typical illustration (Theorem 2 of Rockafellar [51]) is provided by the game-theoretic problem of verifying, for a kernel M(s,t) on S x T, that

$$\min_t \sup_s M(s,t) = \max_s \inf_t M(s,t), \tag{1.6}$$

and that (s*,t*) is one of the saddle points that 1.6 entails. Defining

$$h_2(t) \equiv \sup_s M(s,t) \text{ and } k_2(s) \equiv \inf_t M(s,t),$$

1.3, and hence 1.6, may be established by verifying that the "column sup" $h_2(t^*)$ and "row inf" $k_2(s^*)$ are equal, simultaneously verifying that t* minimizes $h_2(t)$, s* maximizes $k_2(s)$, and (s*,t*) is a saddle point of M.

Another illustration is provided by the treatment in Meeks and Francis [41] of a certain location problem P and a generalized Neyman-Pearson problem D as a dual to P.

So much for the role of weak duality. As to the notion of subgradient or support from convex analysis, its role is in helping identify a certain natural formal Lagrangian version of given problem P, as a problem D weakly dual to P. Indeed the primal-dual problem pair (P^o, D^o) below underlies all of the nonlinear applications we shall treat.

Consider a linear space X, a subset C of X, a (nonlinear) mapping $h: C \to E_1$, and a vector of linear mappings $\{\theta_i\}_{i=1}^n: X \to E_n$. Then, define the region $R \subset X$ by

$$x \in R \iff \begin{cases} \theta_i(x) = c_i, \ 1 \leq i \leq m, \\ \theta_i(x) \geq d_i, \ m+1 \leq i \leq n, \\ x \in C, \end{cases}$$

and consider the

Problem P^o: $\min h(x)$, $x \in R$.

Next define the linear functional $\ell_{h,x_o}(\cdot)$ to be a subgradient of $h(\cdot)$

at x_o relative to R, if $x_o \in R$ and

$$h(x) - h(x_o) \geq \ell_{h,x_o}(x - x_o), \ \forall x \in R. \tag{1.7}$$

(cf. Hendrickson and Buehler [22]). Let L_h denote all such $\ell_{h,x_o}(\cdot)$'s that satisfy $\ell_{h,x_o}(x_o) = h(x_o)$, and for the n-dimensional real vector $y = (y^1, y^2)$, let

$$k(y) = \sum_{1}^{m} y_i^1 c_i + \sum_{m+1}^{n} y_i^2 d_i.$$

Then define the region $S \subset E_n$ by

$$y \in S \Longleftrightarrow \begin{cases} (y^1, y^2) \in E_m \times E_{n-m}^{+}, \text{ and} \\ \\ \text{there is an } \ell_{h,x_o}(\cdot) \in L_h \text{ such that} \\ \\ \ell_{h,x_o}(x) - \sum_{1}^{n} y_i \theta_i(x) \geq 0, \ x \in R, \end{cases}$$

and consider the

Problem D^o: max $k(y)$, $y \in S$.

Now, (P^o, D^o) is a weakly dual pair since, for $x \in R$ and $y \in S$,

$$h(x) \geq h(x_o) + \ell_{h,x_o}(x - x_o) = h(x_o) + \ell_{h,x_o}(x) - \ell_{h,x_o}(x_o) \geq \sum_{1}^{n} y_i \theta_i(x)$$

$\geq k(y)$, and hence all previous remarks (cf. 1.1-1.3) are valid for (P^o, D^o) also. This "constructive primal-Lagrangian" method for solving a problem of type P^o in conjunction with one of type D^o essentially

accounts for our treatment of the nonlinear examples below.

Our first example is an interpretation, in the above vein, of the treatment in Hoeffding [24] of a certain large deviation problem, namely, equating two expressions, one information-theoretic and the other based on the cumulant-generating function, for the large deviation rate of $\Pr_F\{\sum_1^n Z_i \geq 0\}$, where Z_i's are i.i.d. random variables with common probability measure F. Following [24], let F on E_1 have negative expectation and assign some mass to E_1^+; define

$$k_3(y) \equiv - \ln[E_F\{\exp(yZ_1)\}]. \tag{1.8}$$

Let also $\mathcal{G}$ be a class of probability measures G absolutely continuous w.r.t. F, with nonnegative expectation, and define

$$h_3(G) \equiv \int \ln[dG/dF]dG. \tag{1.9}$$

(1.9 is called a Kullback-Leibler information number.)

Then consider the problem pair,

$$P_3: \quad \min h_3(G), \ G \in \mathcal{G},$$

$$D_3: \quad \max k_3(y), \ y \in E_1^+.$$

It may be verified that, using Jensen's inequality as in [24], the pair

(P_3,D_3) is weakly dual, i.e.,

$$h_3(G) \geq k_3(y), \text{ for } G\epsilon\mathcal{G} \text{ and } y\epsilon E_1^+.$$

Furthermore, defining the subclass $\mathcal{G}_o$ of $\mathcal{G}$ by

$$\mathcal{G}_o \equiv \{G_y;\ G_y \text{ s.t. } dG_y/dF(z) = \exp[yz + k_3(y)],\ y\epsilon E_1^+\}, \quad (1.10)$$

and letting y^* be the maximizer in E_1^+ of $k_3(y)$, the solution pair (G_{y^*},y^*) yields $h_3(G_{y^*}) = k_3(y^*)$, and hence yields,

$$\min_{G\epsilon\mathcal{G}} \int \ln[dG/dF]dG = \max_{y\epsilon E_1^+} - \ln[E_F\{\exp(yZ_1)\}], \quad (1.11)$$

the analogue of 1.3, which in turn equalizes the two forms of the large deviation rate.

This treatment of the pair (P_3,D_3) may be viewed in the light of the following treatment of the related pair (P_3^o,D_3^o) as a specialization of the pair (P^o,D^o). (Note that, to fix ideas, but without essential loss of generality, the pair (P_3^o,D_3^o) treats the version of the problem pertaining to probability densities on $(E_1,\mathcal{C},\mu)$.)

(P_3^o,D_3^o) is identified as a specialization of (P^o,D^o) by identifying problem components as follows:

X ~ space of Borel-measurable functions g.

C ~ cone of non-negative Borel-measurable functions g.

For f a probability density,

$h(x) \sim h_3(g) \equiv \int \ln[g/f]gd\mu$.

$(m,n) \sim (1,2)$

$$\theta_1(x) = c_1 \sim \int gd\mu = 1 \tag{1.12a}$$

$$\theta_2(x) \geq d_2 \sim \int zg(z)d\mu(z) \geq 0 \tag{1.12b}$$

R ~ subclass of C that satisfies 1.12a and 1.12b.

$x_o \sim g_o(z) = \exp(y_0 + y_1 z) \cdot f(z)$.

$\ell_{h,x_o}(x) \sim \int \ln[g_o/f]gd\mu$.

S ~ set of $y \equiv (y_0, y_1)$ such that

$$\int \ln[g_o/f]gd\mu - y_0 \int gd\mu - y_1 \int zg(z)d\mu(z) \geq 0, \quad g \in R.$$

Further, the objective-function-equalizing pair may be constructed as follows: Defining

y_1^*: value of y_1 in E_1^+ (guaranteed to exist by external conditions imposed by Hoeffding [24]--see also the "standard condition" of Bahadur [1]) for which $\int y_1^* \exp(y_1^* z) \cdot f(z)d\mu(z) = 0$, and

$$y_0^* = -\ln[\int \exp(y_1^* z) \cdot f(z)d\mu(z)], \tag{1.13}$$

the pair (g^*,y^*) is given by

$$g^*(z) = \exp(y_0^* + y_1^* z) \cdot f(z), \text{ and}$$

$$y^* = (y_0^*, y_1^*).$$

Thus P_3^o and D_3^o have the common extremum given in 1.13. Finally, under the standard condition, the right hand side of 1.13 is, as well, the maximum of the negative of the cumulant generating function of f, establishing the equality of the objective functions of P_3 and D_3, as given by 1.11.

Chapter II (also [32]) treats the Markovian analogue of the above large deviation problem: Let $(S,\mathcal{B},\mu)$ be a finite measure space. Consider a stationary discrete-time Markov process $\{X_i\}$ whose transition probability density kernel $f(y|x)$ is primitive. Also consider a bounded measurable function $a(x,y)$ on S x S satisfying certain regularity conditions given by (A2)-(A4) in Chapter II. (Note that these conditions serve as well to guarantee the existence of certain optimal solutions, in an external sense alluded to in the first paragraph of this chapter.) Next, define

$$k_4(t) \equiv -\ln\lambda_t,$$

where λ_t is the dominant eigenvalue of the kernel

$$K_t(x,y) = \exp[ta(x,y)]f(y|x),\ t\in E_1.$$

Also define the bivariate analogue of the Kullback-Leibler information number

$$h_4(g) \equiv \int\int \ln[g(y|x)/f(y|x)]g(x,y)d\mu^2,$$

where $g(y|x)$ is the conditional probability density kernel corresponding to a bivariate density $g(x,y)$ on S x S w.r.t. μ^2, and define the class $\mathcal{G}$ of densities g that satisfy

$$\int g(x,y)d\mu(y) = \int g(y,x)d\mu(y) \text{ on a.a. } x\in S$$

and

$$\int\int a(x,y)g(x,y)d\mu^2 \geq 0.$$

Then consider a pair

Problem P_4: min $h_4(g)$, $g\in\mathcal{G}$

Problem D_4: max $k_4(t)$, $t\in E_1^+$.

In Chapter II, it is shown that (P_4,D_4) is a weakly dual pair and, moreover, that

$$h_4(g^*) = k_4(t^*),$$

with t^*, the unique minimizer in E_1^+ of λ_t, and $g^*(x,y) \equiv K_{t^*}(x,y)\phi^*(x)\psi^*(y) \cdot \lambda_{t^*}^{-1}$, where $\phi^*(x)$ and $\psi^*(y)$ are the (suitably normalized) left and right eigenfunctions corresponding to λ_{t^*}.

Hence the analogue of 1.3 holds, yielding the two expressions $\min h_4(g)$ and $\max k_4(t)$ for the large deviation rate of $\Pr\{\sum_1^n a(x_{i-1},x_i) \geq 0\}$. Note that, as for the previous i.i.d. example, underlying the weakly dual pair (P_4,D_4) there is a pair (P_4^o,D_4^o) weakly dual in the sense of (P^o,D^o).

Chapter III (also [31]) treats two extremization problems involving mutual information. This study was motivated chiefly by two complementary problems in information theory; the problem of determining (via maximizing mutual information) the capacity of a given channel (viz. conditional distribution kernel) subject to side conditions on the input (viz. marginal distribution), and the problem of determining (via minimizing mutual information) the optimal channel subject to some fidelity criterion stated in conjunction with a fixed input. A certain abstract

representation of these two problems, to be summarized below, not only provides an extension of some previous results (cf. Kolmogorov [35], Berger [8]), but also establishes the saddle point property of a bivariate distribution with respect to mutual information.

Consider a product measure space $(X \times Y, \mathcal{B} \times \mathcal{C}, \mu \times \nu)$. Let $p(x)$, $q(y|x)$, and $r(y)$ be, respectively, a marginal probability density (w.r.t. μ) on X, a conditional probability density (w.r.t. ν) kernel on XxY, and the corresponding marginal probability density on Y, given by $\int q(y|x)p(x)d\mu(x)$.

For the maximization problem, consider a fixed pair $\{p^*(x), q^*(y|x)\}$ and the corresponding marginal density $r^*(y)$, define

$$h_5(p) \equiv \int\int p(x)q^*(y|x) \ln[q^*(y|x)/\int q^*(y|t)p(t)d\mu(t)]d(\mu \times \nu),$$

and the class $\mathcal{P}$ of p;

$$p \in \mathcal{P} \Longleftrightarrow \int\int p(x)q^*(y|x) \ln[q^*(y|x)/r^*(y)]d(\mu \times \nu) = h_5(p^*) < +\infty.$$

Also let $k_5(u) \equiv u \cdot h_5(p^*)$, and define the set U_p;

$$u \in U_p \Longleftrightarrow \begin{cases} u \in E_1, \text{ and} \\ \text{there is an } \ell_{h_5,p^*}(\cdot) \in L_{h_5} \text{ such that} \\ \ell_{h_5,p^*}(p) - u\int\int p(x)q^*(y|x) \ln[q^*(y|x)/r^*(y)]d(\mu \times \nu) \leq 0, \\ p \in \mathcal{P}. \end{cases} \tag{1.14}$$

Then, consider a pair

$$\text{Problem } P_5^o\text{: } \max h_5(p),\ p\in\mathcal{P},$$

$$\text{Problem } D_5^o\text{: } \min k_5(u),\ u\in U_p.$$

That (P_5^o, D_5^o) is a weakly dual pair is immediate, and the choice $p = p^*$, $u = u^* = 1$ trivially yields the analogue of 1.3; $\max_{p\in\mathcal{P}} h_5(p) = \min_{u\in U_p} k_5(u)$. Note that, with this choice (p^*,u^*), 1.14 holds with equality since the linear functional $\ell_{h_5,p^*}(\cdot)$ has the form

$$\ell_{h_5,p^*}(p) = \int\int p(x)q^*(y|x)\ \ln[q^*(y|x)/r^*(y)]d(\mu x\nu).$$

The minimization problem features a certain symmetry with respect to the maximization problem. Define

$$h_6(q) = \int\int p^*(x)q(y|x)\ \ln[q(y|x)/\int q(y|t)p^*(t)d\mu(t)]d(\mu x\nu),$$

and the class $\mathcal{Q}$ of q;

$$q\in\mathcal{Q} \Longleftrightarrow \int\int\{p^*(x)\ \ln[q^*(y|x)/r^*(y)]\}q(y|x)d(\mu x\nu) = h_6(q^*) < +\infty.$$

Also let $k_6(u) \equiv u \cdot h_6(q^*)$, and define the set U_q;

$$u\in U_q \Longleftrightarrow \begin{cases} u\in E_1, \text{ and} \\ \text{there is an } \ell_{h_6,q^*}(\cdot)\in L_{h_6} \text{ such that} \\ \ell_{h_6,q^*}(q) - u\int\int\{p^*(x)\ \ln[q^*(y|x)/r^*(y)]\}q(y|x)d(\mu x\nu)\geq 0, \end{cases}$$

ℓ for $q \in \mathcal{Q}$.

Then, for the pair

Problem P_6^o: $\min h_6(q)$, $q \in \mathcal{Q}$,

Problem D_6^o: $\max k_6(u)$, $u \in U_q$,

entirely analogously to the treatment to the pair (P_5^o, D_5^o), we can establish that $\min_{q \in \mathcal{Q}} h_6(q) = \max_{u \in U_q} k_6(u)$.

The final remark pertaining to Chapter III is that, in addition to the fact that the actual presentation in Chapter III is given in terms of probability measures rather than densities, the approach actually used there features problem pairs (P_i, D_i) equivalent to, but slightly more direct, than the primal-Lagrangian pairs (P_i^o, D_i^o) of the above presentation.

Chapter IV (to which the last remark in fact pertains as well) considers the easy (noncharacterizing) direction of Lanford-Ruelle type theorem, for not necessarily stationary processes. Specifically, we show in elementary fashion, applying the above primal-Lagrangian method, that a certain stationary Markov process minimizes the specific free

energy, among processes f possessing densities f_n on the products $(S^n, \mathcal{B}^n, \nu^n)$ of a finite measure space $(S, \mathcal{B}, \nu)$.

The first step of our argument is to show that any f^* with

$$f_n^* = C_n \exp\{-\sum_1^{n-1} U(x_i, x_{i+1})\} \text{ minimizes}$$

$$h_7^n(f) \equiv \int\{\sum_1^{n-1} U(x_i, x_{i+1}) + \ln f_n\} f_n d\nu^n,$$

where $U(x,y)$ is ν^2-integrable. For this, letting K^+ be the cone of nonnegative measurable functions on S^n, and defining the class $\mathfrak{F}$ by

$$f \in \mathfrak{F} \Longleftrightarrow \begin{cases} \int f_n d\nu^n = 1, \text{ and} \\ f_n \in K^+, \end{cases}$$

consider

Problem P_7^o: min $h_7^n(f)$, $f \in \mathfrak{F}$.

Also defining the set U by

$$u \in U \Longleftrightarrow \begin{cases} u \in E_1, \text{ and} \\ \text{there is an } \ell_{h_7, f^*(\cdot)} \in L_{h_7} \text{ such that} \\ \ell_{h_7, f^*}(f_n) - u\int f_n d\nu^n \geq 0 \text{ for } f \in \mathfrak{F}, \end{cases}$$

consider

Problem D_7^o: max u, $u \in U$.

Then (P_7^o, D_7^o) is a weakly dual pair, and the choice $f = f^*$, together with the choice $u^* = \ln C_n$ yields the analogue of 1.2; $h_7^n(f^*) = u^*$, and hence the analogue of 1.3;

$$\min_{f \in \mathfrak{F}} h_7^n(f) = \max_{u \in U} u.$$

In the second step, $h_7^n(f^*)$ is to be compared with $h_7^5(f^{**})$, where f^{**} is a certain Markov process with marginal densities

$$f_n^{**} \equiv \phi(x_1)\psi(x_n) \exp[-\sum_{1}^{n-1} U(x_i, x_{i+1})]/\lambda^{n-1},$$

detailed in Chapter IV. Then finally, based on the argument of the above two steps, we reach the following Lanford-Ruelle analogue;

$$\text{for } f \in \mathfrak{F}, \ \underline{\lim}_n \, n^{-1} h_7^n(f) \geq \lim_n n^{-1} h_7^n(f^{**}) = -\ln \lambda.$$

II. LARGE DEVIATIONS FOR MARKOV PROCESSES

1. Introduction

Koopmans [36] has considered the rates of decay of probabilities of "nonlocal" (in the sense of Chernoff [10]) errors of sequences of likelihood ratio tests discriminating between two Markov processes P and Q. In Koopmans' setting, where P and Q are discrete-time and stationary, these rates of decay are expressible in terms of a certain extremal dominant root. Koopmans' work in effect determines the rate of decay of the probabilities of large deviations of the sample averages of the function $\ln[q(x_i|x_{i-1})/p(x_i|x_{i-1})]$ of observed transitions; i.e., of the probabilities of events

$$E_n: \ \{\sum_{i=1}^{n} \ln[q(X_i|X_{i-1})/p(X_i|X_{i-1})] \geq 0\}.$$

Slight modification of the argument in [36] verifies that the rate of decay of the probabilities of events

$$F_n: \ \{\sum_{i=1}^{n} a\,(X_{i-1},X_i) \geq 0\},$$

$a(x,y)$ not necessarily of the form $\ln[q(y|x)/p(y|x)]$, similarly is expressible in terms of an analogous extremal dominant root. (As pointed

out by Bahadur [2], one may establish the asymptotic equivalence of the forms $a(x,y)$ and $\ln[q(y|x)/p(y|x)]$ based on the transformation of our Remark 2, analogously to the i.i.d. case of Bahadur and Raghavachari [3]. However, our choice $a(x,y)$ appeared to be natural to the setting of Harris [21] adopted below, and also facilitated our establishing part (iii) of Theorem 2.3.1.)

A related investigation, by Boza [9], dealing with the comparison of tests of hypotheses concerning finite state space stationary Markov chains, involves the rate of decay of probabilities of certain events G_n defined in terms of transition counts; specializations of these events G_n are of the form F_n. Boza found the rate of decay of the probabilities of the events G_n (and hence of the events F_n) to be given by a certain extremal information functional.

Thus two expressions, one function-analytic or spectral [36] and the other information theoretic [9], are available for the rate of decay of the probabilities of events F_n, with the second restricted to finite state space.

Also available in the literature are corresponding "function-analytic" (Chernoff [10], Bahadur and Rao [4]) and information theoretic (Sanov [54]) rates of decay in the analogous i.i.d. case (with the validity of neither restricted to the finite case), and reconciliations of the two, essentially using mathematical programming duality, by Hoeffding [24] and Whittle [62], as sketched in Chapter I.

This study, expanding on [33], brings the duality approach to the Markov case. Our objective here is not only to extend to this case the "direct" duality point of view of [24] and [62]; it is, in addition, to extend the work in [9] to the not-necessarily-finite case. This roundabout approach to extending [9] (by way of [36] plus duality) may not be entirely unnatural, since the argument in [9] is combinatoric.

Assumptions are laid out in section 2. In section 3, some properties of the relevant dominant root λ_t are examined, in a manner analogous to Koopmans'. Among these is an expression for the derivative λ_t' of λ_t, of use in the later development. We also compute the rate of decay of the probabilities of the event F_n in terms of λ_t, in a manner entirely

analogous to that in [36]. Section 4 implements the program of the previous paragraph; the point of view here is essentially Hoeffding's, extended to the Markov case and viewed in the light of mathematical programming duality. Finally, section 5 is set aside for a supplementary note, outlining the proof of the uniform convergence of $\phi_{t+\delta}$ to ϕ_t, to support the argument used for part (iii) of Theorem 2.3.1 below.

2. Assumptions

Let $(S,\mathcal{B},\mu)$ be a finite measure space. Following Harris [21], consider a transition probability density kernel $p(y|x)$ that (a) is $\mathcal{B} \times \mathcal{B}$ - measurable, and (b) has the property that there are real numbers c and d, and a positive integer N, such that the N'th iterate $p^{(N)}(y|x)$ of $p(y|x)$ satisfies

$$0 < c \leq p^{(N)}(y|x) \leq d < +\infty. \tag{A1}$$

By Theorem 10.1 of [21], the kernel $p(y|x)$ has a bounded left eigenfunction $\phi(x)$ corresponding to the dominant root 1.

Now consider a $\mathcal{B} \times \mathcal{B}$ - measurable function $a(x,y)$ satisfying

$$|a(x,y)| < M < +\infty, \text{ on } S \times S. \tag{A2}$$

$$\iint a(x,y)p(y|x)\,\phi(x)\,d\mu^2 < 0, \tag{A3}$$

and, for some $\varepsilon > 0$,

$$\operatorname*{ess\,inf}_{x} \left[\int_{\{a(x,y)>\varepsilon\}} p(y|x)\,d\mu(y)\right] > 0. \tag{A4}$$

Also define kernels $K_t(x,y) = e^{ta(x,y)}p(y|x)$, $-\infty < t < +\infty$. In view of (A1) and (A2), K_t satisfies as well the assumptions of Theorem 10.1 of [21] for all t, so that K_t possesses a positive eigenvalue λ_t, and corresponding left and right eigenfunctions ϕ_t and ψ_t which are henceforth normalized so that $\int \phi_t \psi_t d\mu = 1$. For $(x,y)\in S\times S$,

$$K_t^{(n)}(x,y) = \lambda_t^n \psi_t(x)\,\phi_t(y)[1 + g_t(x,y;n)], \quad \forall t, \tag{2.2.1}$$

where $|g_t(x,y;n)| \leq \Delta_t^n$, with $0 < \Delta_t < 1$.

Some remarks on assumptions (A3) and (A4) also are in order; (A3) is used below in conjunction with (A2) to guarantee that λ_t is decreasing near zero, while (A4) is used to guarantee that λ_t tends to $+\infty$ with t. In the i.i.d. case, analogous conditions have been used by Bahadur [1] to ensure him the "standard condition".

3. The decay rate and the dominant root λ_t

Let $\{X_i\}_{i=0}^{\infty}$ form a Markov process over state space S with initial

density $p_0(x)$ and transition density $p(y|x)$. Let $M_n(t)$ be the moment generating function of $S_n \equiv \sum_{i=1}^{n} a(X_{i-1},X_i)$.

Theorem 2.3.1: Under assumptions (A1) and (A2),

(i) $\lim_n [M_n(t)]^{1/n} = \lambda_t$, $\forall t$.

(ii) λ_t is convex and analytic, $\forall t$.

(iii) $\lambda_t' = \int\int a(x,y)K_t(x,y)\phi_t(x)\psi_t(y)d\mu^2$, $\forall t$.

Under assumptions (A1)-(A4),

(iv) λ_t achieves a unique minimum at t^*, $0 < t^* < +\infty$.

Proof: (i) We shall in fact prove the slightly stronger

$$\lim_n [M_n(t)/\lambda_t^n] = A_t, \tag{2.3.1}$$

where $A_t = \int\int \psi_t(x)\phi_t(y)p_0(x)d\mu^2 > 0$. Now $M_n(t) = \int\int K_t^{(n)}(x,y)p_0(x)d\mu^2$ $= \lambda_t^n(A_t + B_{t,n})$, with $B_{t,n} = \int\int \psi_t(x)\phi_t(y)g_t(x,y;n)p_0(x)d\mu^2$, where the first equality is by definition, and the second holds in view of 2.2.1, which also entails $\lim_n B_{t,n} = 0$.

(ii) Convexity follows from part (i), and the fact that $M_n(t)^{1/n}$ is convex in t for all n. As for analyticity, following Koopmans [36] consider the bilateral Laplace transform $H_n(z)$ of $F_n(x) \equiv P_r\{S_n \leq x\}$,

and, for any $r > 0$, let $T_r = \{t:|t| \leq r\}$. In view of (A2), $|[H_n(z)]^{1/n}| \leq e^{Mr}$ for every z with real part in T_r, and for every $n = 1,2,\ldots$. Hence each $[H_n(z)]^{1/n}$ is analytic in the infinite strip $\mathfrak{I}_r:\{z = t + iu,\ t\epsilon T_r\}$. Thus part (i), the uniform bound e^{Mr} for $|[H_n(z)]^{1/n}|$, and Vitali's theorem imply that $\lim_n[H_n(z)]^{1/n}$ is analytic in the interior of $\mathfrak{I}_r$; hence that $\lim_n[H_n(t)]^{1/n} = \lim_n[M_n(t)]^{1/n} = \lambda_t$ is analytic in the interior of T_r.

(iii) Let $\delta > 0$, and consider

$$(\lambda_{t+\delta} - \lambda_t)\psi_{t+\delta}(x) + \lambda_t[\psi_{t+\delta}(x) - \psi_t(x)] \tag{2.3.2}$$

$$= \lambda_{t+\delta}\psi_{t+\delta}(x) - \lambda_t\psi_t(x)$$

$$= \int K_{t+\delta}(x,y)[\psi_{t+\delta}(y) - \psi_t(y)]d\mu(y) + \int[e^{\delta a(x,y)} - 1]K_t(x,y)\psi_t(y)d\mu(y), \tag{2.3.3}$$

where the second equality follows from subtracting and adding $\int K_{t+\delta}(x,y)\psi_t(y)d\mu(y)$. Now, multiplying 2.3.2 and 2.3.3 by $\phi_{t+\delta}(x)$, integrating w.r.t. x (using Fubini's Theorem for the first addend of 2.3.3), and rearranging, $(\lambda_{t+\delta} - \lambda_t)\{1 - \int[\psi_{t+\delta}(x) - \psi_t(x)]\phi_{t+\delta}(x)d\mu(x)\}$ $= \int\int(e^{\delta a(x,y)} - 1)K_t(x,y)\phi_{t+\delta}(x)\psi_t(y)d\mu^2$. Rearranging

again and dividing by δ,

$$(\lambda_{t+\delta} - \lambda_t)\delta^{-1} = \int\int (e^{\delta a(x,y)} - 1)\delta^{-1} k_\delta(x,y)\, d\mu^2, \qquad (2.3.4)$$

where $k_\delta(x,y) \equiv K_t(x,y)\phi_{t+\delta}(x)\psi_t(y)[\int \phi_{t+\delta}(x)\psi_t(x)\, d\mu(x)]^{-1}$. Similarly,

$$(\lambda_t - \lambda_{t-\delta})\delta^{-1} = \int\int (1 - e^{-\delta a(x,y)})\delta^{-1} k_{-\delta}(x,y)\, d\mu^2. \qquad (2.3.5)$$

Now $\lambda_t' = \lim_{\delta\to 0}(\lambda_{t+\delta} - \lambda_t)\delta^{-1} \geq \lim_{\delta\to 0}\int\int a(x,y)k_\delta(x,y)\, d\mu^2 = \int\int a(x,y)k_0(x,y)\, d\mu^2$.

Here the first equality is due to the fact that λ_t is analytic.

The second equality follows from the uniform (in (x,y)) convergence of k_δ to k_0, since $a(x,y)$ is bounded and $\mu(S)$ is finite; this uniform convergence of k_δ in turn follows from the uniform (in x) convergence of $\phi_{t+\delta}$ to ϕ_t, since the other components of k_δ do not depend on δ; finally, the uniform convergence of $\phi_{t+\delta}$ follows from a straightforward application of the arguments in Theorem 2.4.2 of Conn [13] or Theorem 5.2 of Madsen and Conn [40], with $(K_t^{(N)}, K_{t+\delta}^{(N)})$, $|\delta| < 1/M$, replacing (M_n, M_{n+1}), as sketched in section 5.

The inequality follows from 2.3.4, from the fact that the limit exists, and the fact that $e^{\delta a} - 1 \geq \delta a$.

Similarly, using 2.3.5 and the comparison $1 - e^{-\delta a} \leq \delta a$,

$$\lambda_t' \leq \iint a(x,y)k_0(x,y)d\mu^2.$$

(iv) (A3) and part (iii) imply that

$$\lambda_0' = \iint a(x,y)p(y|x)\phi_0(x)d\mu^2 < 0. \tag{2.3.6}$$

Now, by (A4), $\exists\ \eta > 0$ ∋ $\operatorname{ess\,inf}_{x\ \{a(x,y) > \varepsilon\}} [\int p(y|x)d\mu(y)] \geq \eta$. Hence assuming $\int \phi_t d\mu = 1$,

$$\lambda_t = \iint e^{ta(x,y)}p(y|x)\phi_t(x)d\mu^2 \geq \int [\int_{\{a(x,y) > \varepsilon\}} e^{ta(x,y)}p(y|x)d\mu(y)]$$

$$\phi_t(x)d\mu(x) \geq e^{t\varepsilon} \cdot \eta \to +\infty \text{ as } t \to +\infty. \tag{2.3.7}$$

2.3.6, 2.3.7, and convexity assure that λ_t achieves its minimum in $(0,+\infty)$, the uniqueness of which is guaranteed by the fact that λ_t, being analytic for all t, cannot be flat in a proper subinterval without being flat everywhere.

The remaining part of this section is to verify that $\ln \lambda_{t*}$ is the relevant large deviation rate. Although the argument used below to this effect is parallel to that of Koopmans', we state this fact as a theorem and sketch the proof for the sake of completeness.

Theorem 2.3.2 (essentially Koopmans' Theorem 3): Under the assumptions (A1)-(A4),

$$\lim_n \Pr\{S_n \geq 0\}^{1/n} = \lambda_{t^*} \tag{2.3.8}$$

Proof: First of all, note that $\Pr\{S_n \geq 0\} \leq E\{\exp(tS_n)\} = M_n(t)$, for $t \geq 0$. Now, (i) of Theorem 2.3.1 yields $\overline{\lim_n} \Pr\{S_n \geq 0\}^{1/n} \leq \lambda_{t^*}$. Next, note that, in view of (ii) and (iv) of Theorem 2.3.1, for any $t_o > t^*$ and $b > 0$, there is an $s > 0$ such that

$$\lambda_{t_o - s}/\lambda_{t_o} < 1 \quad \text{and} \quad \lambda_{t_o + s}/\lambda_{t_o} < e^{bs} \tag{2.3.9}$$

(For the latter, one may observe that convexity of λ_t implies $\lambda_{t_o+s}/\lambda_{t_o} < 1 + (\lambda'_{t_o+s}/\lambda_{t_o})\cdot s$). Using estimates of the probability of various sets, namely,

$$\Pr\{0 \leq S_n/n \leq b\} \leq e^{nbt}\Pr\{S_n \geq 0\}/M_n(t),$$

$$\Pr\{S_n < 0\} \leq M_n(t - s)/M_n(t), \text{ and}$$

$$\Pr\{S_n/n > b\} \leq e^{-nbs}M_n(t + s)/M_n(t),$$

we have

$$\Pr\{S_n \geq 0\}^{1/n} \geq e^{-bt}\{M_n(t)[1 - M_n(t - s)/M_n(t) - e^{-nbs}M_n(t + s)/M_n(t)]\}^{1/n} \tag{2.3.10}$$

for all $b > 0$, $s > 0$, and $t > 0$.

Now, in view of 2.3.1 and 2.3.9, the right hand side of 2.3.10 with $t = t_o$ tends to $e^{-bt_o}\lambda_{t_o}$ with n, so that $\underline{\lim}_n \Pr\{S_n \geq 0\}^{1/n} \geq e^{-bt_o}\lambda_{t_o}$. But $b > 0$ was arbitrary, so that $\underline{\lim}_n \Pr\{S_n \geq 0\}^{1/n} \geq \lambda_{t_o} \geq \lambda_{t^*}$.

4. The two mutually weakly dual problems

As indicated at the end of section 1, this section brings the direct duality-related point of view to our Markov setting, by equating the decay rate $\ln \lambda_{t^*}$ of the previous section to a not-necessarily-finite analogue of Boza's information theoretic decay rate. This equality, combined with 2.3.8, extends the validity of Boza's rate expression.

For a bivariate probability density f(x,y) on SxS, let h(x) and g(y|x) be the marginal probability density on S and the essentially unique conditional probability density kernel on SxS, respectively. Now let the set $\mathfrak{F}_a$ of densities f satisfy

$$\int f(x,y)d\mu(y) = \int f(y,x)d\mu(y), \text{ almost all } x, \tag{2.4.1}$$

$$\iint a(x,y)f(x,y)d\mu^2 \geq 0, \tag{2.4.2}$$

and

$$\iint \ln[g(y|x)/p(y|x)]f(x,y)d\mu^2 < +\infty. \tag{2.4.3}$$

Define $I(f,p) \equiv \iint \ln[g(y|x)/p(y|x)]f(x,y)d\mu^2$, $f \in \mathfrak{F}_a$.

Theorem 2.4.1: Under assumptions (A1)-(A4), for $f \in \mathfrak{F}_a$ and $t \in (0,+\infty)$,

$$I(f,p) \geq -\ln \lambda_t. \tag{2.4.4}$$

Proof: Note first that 2.4.1 implies that, for any bounded measurable function $s(\cdot)$,

$$\int\int s(x)f(x,y)d\mu^2 = \int\int s(y)f(x,y)d\mu^2. \tag{2.4.5}$$

Now write

$$I(f,p) \geq \int\int \ln[g(y|x)/p(y|x)]g(y|x)h(x)d\mu^2 \tag{2.4.6a}$$

$$-\int\int \ln[\psi_t(y)/\psi_t(x)]g(y|x)h(x)d\mu^2 \tag{2.4.6b}$$

$$-t\int\int a(x,y)g(y|x)h(x)d\mu^2 \tag{2.4.6c}$$

$$= \int\int -\ln[p(y|x)/g(y|x)\cdot\psi_t(y)/\psi_t(x)\cdot e^{ta(x,y)}]g(y|x)h(x)d\mu^2$$

$$\geq -\ln\int\int \psi_t(y)/\psi_t(x)\cdot e^{ta(x,y)}p(y|x)h(x)d\mu^2 \tag{2.4.6d}$$

$$= -\ln\lambda_t.$$

The first inequality is due to the fact that 2.4.6b is zero in view of 2.4.5, and 2.4.6c is nonpositive in view of 2.4.2. The second inequality is Jensen's, and the last equality follows by integrating first w.r.t. y and then w.r.t. x, and appealing to definitions of ψ_t and λ_t given in section 2.

Now let t^* be the unique minimizer of λ_t identified in (iv) of Theorem 2.3.1, and define

$$h^*(x) = \psi_{t^*}(x)\,\phi_{t^*}(x)$$

$$g^*(y|x) = K_{t^*}(x,y)\psi_{t^*}(y)/\lambda_{t^*}\psi_{t^*}(x) \tag{2.4.7}$$

$$f^*(x,y) = h^*(x)g^*(y|x).$$

Corollary 2.4.1: $\min_{f\in\mathfrak{F}_a} I(f,p) = I(f^*,p) = -\ln \lambda_{t^*} = \max_{t\in(0,\infty)} [-\ln \lambda_t]$.

Proof: In view of Theorem 2.4.1, it is sufficient to verify that $f^*\in\mathfrak{F}_a$ and that $I(f^*,p) = -\ln \lambda_{t^*}$, i.e., that equality holds in the statement of Theorem 2.4.1 for $f = f^*$ and $t = t^*$ (cf. Relations (1.1)-(1.3) in Chapter I). That $f^*\in\mathfrak{F}_a$ is easily verified. That equality holds in Theorem 2.4.1 for $f = f^*$ and $t = t^*$ follows from verifying equality in 2.4.6a and 2.4.6d. Regarding 2.4.6a, one needs to verify only that expression 2.4.6c equals 0 when $f = f^*$, which follows from (iii) and (iv) of Theorem 2.3.1 since 2.4.6c equals $(-t^*\lambda_{t^*}^{-1})\int\int a(x,y)K_{t^*}(x,y)\phi_{t^*}(x)\psi_{t^*}(y)d\mu^2 = (-t^*\lambda_{t^*}^{-1})\lambda'_{t^*} = 0$. Regarding 2.4.6d, the argument of ln (·), namely $[e^{t^*a(x,y)}p(y|x)\psi_{t^*}(y)/(g^*(y|x)\psi_{t^*}(x))]$, equals the constant λ_{t^*}, in view of 2.4.7.

Remark 1: As suggested by Hoeffding [24], mutually dual problem pairs typically admit certain irregular parametric cases for which D is "unbounded" and P is "infeasible". (See P_4 and D_4 in Chapter I for P and D.) This feature is present also here, if (A4) is replaced by (A4)':

$\int p(y|x)\,d\mu(y) = 1$, for all $x \in S$. In this case, $(a)\mathfrak{F}_a$ is empty, and $\{a(x,y) < -\varepsilon\}$

(b) $\sup_{t\in(0,\infty)} \ln \lambda_t^{-1} = +\infty$.

To see (a), note that (A4)' implies that $\int\int a(x,y)p(y|x)h(x)\,d\mu^2 < 0$, for any $h(x)$, which, together with 2.4.3, implies that $\int\int a(x,y)g(y|x)h(x)\,d\mu^2 < 0$, for any g. To show (b), observe that (A4)' implies that $\lambda_t = \int\int e^{ta(x,y)} p(y|x)\,\phi_t(x)\,d\mu^2 < e^{-t\varepsilon}$, $t > 0$.

The symmetric alternative, with unbounded P and infeasible D, cannot be exhibited in view of nonnegativity of I(f,p), i.e., P admits a natural lower bound. However, the parametric case where minimization of I(f,p) achieves the lower bound 0 can be tied to the case where $\lambda_0' \geq 0$. For suppose that (A3) is replaced by (A3)': $\int\int a(x,y)p(y|x)\,\phi(x)\,d\mu^2 \geq 0$.

Then $\min_{f\in\mathfrak{F}_a} I(f,p) = 0$, trivially with $f(x,y) = p(y|x)\,\phi_0(x)$, and $\sup_{t\in(0,\infty)} \ln \lambda_t^{-1} = \ln \lambda_0^{-1} = 0$.

Remark 2: The parametrized family

$$\mathfrak{F}^0 = \{f_t;\ f_t(x,y) = K_t(x,y)\phi_t(x)\psi_t(y) \cdot \lambda_t^{-1},\ t \geq 0\}$$

is a Markov analogue of a construct that has repeatedly been used for the i.i.d. case, for example by Cramér [14], Khinchin [29], Chernoff [10], [11], Bahadur [1], Bahadur and Raghavachari [3], and Feller [19]. Now, for the additionally constrained optimization problem, i.e.,

$$\text{minimize } I(f_t,P) \text{ over } \mathfrak{F}^0,$$

the verifying of the weak duality is extremely simple and, in doing so, one may identify the feature that underlies a well-known necessary condition for optimality. In fact, by a straight forward computation, we have,

$$I(f_t, p) = t \cdot \rho'(t) - [\rho(t) - \rho(0)],\ t \geq 0,$$

where $\rho(t) = \ln \lambda_t$. (Hence $\rho(0) = 0$). But, since

$$t \cdot \rho'(t) = t \cdot \lambda_t'/\lambda_t = t\iint a(x,y)f_t(x,y)d\mu^2 \cdot \lambda_t^{-1} \geq 0,$$

$I(f_t,P) \geq -\rho(t)$, i.e., the weak duality has been verified.

To establish that

$$t \cdot \rho'(t) = 0, \text{ for some } t \geq 0, \tag{2.4.8}$$

(hence to establish the duality), we must choose $t^* > 0$ such that $\rho'(t^*) = 0$, or choose $t^* = 0$ if $\rho'(t) > 0$, $\forall t \geq 0$, (cf. end of Remark 1). This condition that the inner product 2.4.8 becomes zero under an optimal t^* is known as "complementary slackness" in the mathematical programming lingo. This particular inner product 2.4.8 has a certain bearing on the duality between the natural parameter space and the expectation space of the exponential family, in the sense (Efron [18]) that t is considered as a parameter value and $\rho'(t)$ an expectation generator.

Remark 3: The restriction 2.4.1 is in fact an ergodicity-related condition, since it is equivalent to the condition that h is an eigenfunction of g corresponding to the root 1.

5. Uniform convergence of $\phi_{t+\delta}$ to ϕ_t

We outline the proof of the fact that $\phi_{t+\delta} \to \phi_t$ uniformly as $\delta \to 0$, in the spirit of Conn [13] and Madsen and Conn [40]. Facts (1)-(6) below are preliminaries.

(1) Bounds for $K^{(N)}_{t+\delta}$, $\phi_{t+\delta}$, and $\lambda_{t+\delta}$ may be computed in elementary fashion, where $|\delta| < 1/M$: Let $E(n) \equiv \exp[nN(|t|M + 1)]$, n an integer.

(1.i) $E(-1)c \le K^{(N)}_{t+\delta}(x_0,x_N) \le E(1)d$. Since

$$K^{(N)}_{t+\delta}(x_0,x_N) = \int\cdots(N-1)\cdots\int \exp[(t+\delta)\sum_{i=1}^{N} a(x_{i-1},x_i)] \prod_{i=1}^{N} p(x_i|x_{i-1})d\mu^{N-1}$$

$$\le \exp[N(|t|M+1)]p^{(N)}(x_N|x_0)$$

$$\le \exp[N(|t|M+1)]d,$$

in view of (A1) and (A2) in section 2, and similarly for the lower bound.

Remark 4: Note that the bounds (1.i) also work for $\lambda^N_{t+\delta}\,\phi_{t+\delta}(x_N) = \int K^{(N)}_{t+\delta}(x_0,x_N)\,\phi_{t+\delta}(x_0)d\mu(x_0)$, assuming the normalization $\int \phi_{t+\delta}d\mu = 1$.

(1.ii) $\mu(S)E(-1)c \le \lambda^N_{t+\delta} \le \mu(S)E(1)d$. Integrate both bounds for $\lambda^N_{t+\delta}\,\phi_{t+\delta}(x_N)$.

(1.iii) $\mu(S)E(-2)c/d \le \phi_{t+\delta}(x_N) \le \mu(S)E(2)d/c$. Divide the lower bound for $\lambda^N_{t+\delta}\,\phi_{t+\delta}(x_N)$ by the upper bound for $\lambda^N_{t+\delta}$ to get the lower bound, and analogously for the upper bound.

(1.iv) $\mu(S)^{-1}E(-4)[c/d]^2 \le K_{t+\delta}(x_0,x_N)\,\phi_{t+\delta}(x_0)[\lambda_{t+\delta}\,\phi_{t+\delta}(x_N)]^{-1} \le \mu(S)^{-1}E(4)[d/c]^2$. Use (1.i), (1.iii), and Remark 4.

(2) $\int|K^{(N)}_{t+\delta}(x_0,x_N) - K^{(N)}_t(x_0,x_N)|d\mu(x_0) \to 0$ uniformly in x_N as $\delta\to 0$.

Since

$$|K^{(N)}_{t+\delta}(x_0,x_N) - K^{(N)}_t(x_0,x_N)|$$

$$\leq \int\cdots(N-1)\cdots\int\left|\exp\left[\delta\sum_{i=1}^{N} a(x_{i-1},x_i)-1\right]\right|\left|\exp\left[t\sum_{i=1}^{N} a(x_{i-1},x_i)\right]\right|\prod_{i}^{N} p(x_i|x_{i-1})\,d\mu^{N-1}$$

$$\leq |e^N - 1|\exp[tNM]p^{(N)}(x_N|x_0) \leq d|e^N - 1|\exp(tNM),$$

apply Lebesgue's bounded convergence theorem.

(3) $|\lambda_{t+\delta} - \lambda_t| \to 0$ as $\delta \to 0$, in view of analyticity.

(4) $\int|K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta} - K^{(Nk)}_t(x_0,x_N)/\lambda^{Nk}_t|\,d\mu(x_0) \to 0$

uniformly in x_N as $\delta \to 0$, for all multiples Nk of N. Analogously to the argument in [13] and [40], this is shown by induction on k, using (1.ii), (2), and (3).

(5) $\left|\int\phi_t(x_N) - \int\phi_t(x_0)K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}d\mu(x_0)\right| \to 0$

uniformly in x_N as $\delta \to 0$ for all multiples Nk of N. Analogously to the argument in [13] and [40], this is shown using (1.iii), (2), and (4).

(6) $|K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta} - \psi_{t+\delta}(x_0)\phi_{t+\delta}(x_N)|$

$$\leq (E(8) - 1)\cdot[1 - E(-4)(c/d)^2]^{Nk-1}$$

This is shown analogously to the argument in lemma 2.1.3 of [13], using (1.iii) and (1.iv).

Finally, note that

$$|\phi_{t+\delta}(x_N) - \phi_t(x_N)| \leq |\phi_{t+\delta}(x_N) - [\int\phi_t(x)\psi_{t+\delta}(x_0)d\mu(x_0)]\phi_{t+\delta}(x_N)|$$

$$+ \left|[\phi_t(x_0)\psi_{t+\delta}(x_0)d\mu(x_0)]\phi_{t+\delta}(x_N) - \phi_t(x_N)\right|$$

$$= \left|A_\delta(x_N)\right| + \phi_{t+\delta}(x_N)\left|\int A_\delta(x_N)d\mu(x_N)\right|, \quad (2.5.1)$$

where $A_\delta(x_N) = \phi_{t+\delta}(x_N) - [\int\phi_t(x_0)\psi_{t+\delta}(x_0)d\mu(x_0)]\phi_{t+\delta}(x_N)$.

Now

$$\left|A_\delta(x_N)\right| \leq \left|\phi_t(x_N) - \int\phi_t(x_0)[K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta}]d\mu(x_0)\right|$$

$$+ \left|\int\phi_t(x_0)[K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta}]d\mu(x_0) - \int\phi_t(x_0)\psi_{t+\delta}(x_0)\phi_{t+\delta}(x_N)d\mu(x_0)\right|$$

$$= \left|\phi_t(x_N) - \int\phi_t(x_0)[K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta}]d\mu(x_0)\right|$$

$$+ \left|\int\phi_t(x_0)[K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta} - \psi_{t+\delta}(x_0)\phi_{t+\delta}(x_N)]d\mu(x_0)\right|$$

$$\leq \left|\phi_t(x_N) - \int\phi_t(x_0)[K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta}]d\mu(x_0)\right| \quad (2.5.2)$$

$$+ \mu(S)^{-1}E(2)(d/d\delta)\left|K^{(Nk)}_{t+\delta}(x_0,x_N)/\lambda^{Nk}_{t+\delta} - \psi_{t+\delta}(x_0)\phi_{t+\delta}(x_N)\right|d\mu(x_0), \quad (2.5.3)$$

where the last expression is uniformly bounded, in view of (1.i), (1.ii), (1.iii), and (6). Hence $A_\delta(x_N)$ is uniformly bounded, and, in view of (1.iii) and Lebesgue's bounded convergence theorem, it is sufficient, for 2.5.1 tending to 0 uniformly in x_N as $\delta \to 0$, to show that $A_\delta(x_N) \to 0$ uniformly in x_N as $\delta \to 0$. But 2.5.3 tends to 0 uniformly in δ as k gets large, in view of (6), and 2.5.2 tending to 0 uniformly in x_N as $\delta \to 0$ for all k, in view of (5).

III. BIVARIATE DISTRIBUTIONS AS SADDLE POINTS OF MUTUAL INFORMATION

1. Introduction

We have observed that the known extremal properties of the multivariate normal and other distributions with regard to the mutual information functional (Kolmogorov [35]) can be viewed in the light of the fact that essentially any probability distribution function F on a Cartesian product X x Y is a saddle point of this functional in a certain sense, under a pair of moment conditions determined by F. This saddle point property is indicated by, and essentially arises from, the concavity-convexity facts of the situation (Lindley [39]), while the specific nature of the two moment conditions is suggested by the variational expression $dI(\theta)/d\theta$ on page 198 of Balakrishnan [6].

Our approach in essence is this: Fix a probability distribution function F on X x Y, considered as a pair $(\alpha,\mathfrak{F})$ whose first coordinate is a marginal distribution function $\alpha(x)$ on X, and whose second coordinate is a kernel of conditional distribution functions $F_x(y)$ on Y. Consider as well an analogous pair $(\beta,\mathfrak{G})$, and define the four-argument functional

$$J(\alpha,\mathfrak{F},\beta,\mathfrak{G}) = \int_X [\int_Y \ln[dF_x/dF_\alpha(y)]dG_x(y)]d\beta(x), \tag{3.1.1}$$

where $F_\alpha(y)$ is the "marginal" distribution function on Y corresponding to the marginal distribution function $\alpha(x)$ on X and the conditional distribution functions $F_x(y)$ on Y, and where $dF_x/dF_\alpha(y)$ is the density of F_x w.r.t. F_α. Note that 3.1.1 is the ordinary mutual information when the second pair of arguments coincides with the first.

We now observe (section 3) that, for all β satisfying

$$J(\alpha,\mathfrak{F},\beta,\mathfrak{F}) = J(\alpha,\mathfrak{F},\alpha,\mathfrak{F}) < +\infty \tag{3.1.2}$$

one has

$$J(\beta,\mathfrak{F},\beta,\mathfrak{F}) \leq J(\alpha,\mathfrak{F},\alpha,\mathfrak{F}), \tag{3.1.3}$$

which fact has already been noted in Kerridge [28] for the finite case.

Also (section 4), for all $\mathfrak{G}$ satisfying

$$J(\alpha,\mathfrak{F},\alpha,\mathfrak{G}) = J(\alpha,\mathfrak{F},\alpha,\mathfrak{F}) < +\infty, \tag{3.1.4}$$

one has

$$J(\alpha,\mathfrak{G},\alpha,\mathfrak{G}) \geq J(\alpha,\mathfrak{F},\alpha,\mathfrak{F}); \tag{3.1.5}$$

in other words, for all $(\beta,\mathfrak{G})$ satisfying the two conditions 3.1.2 and 3.1.4,

$$J(\beta,\mathfrak{F},\beta,\mathfrak{F}) \leq J(\alpha,\mathfrak{F},\alpha,\mathfrak{F}) \leq J(\alpha,\mathfrak{G},\alpha,\mathfrak{G}),$$

which is the saddle point property of the title of this chapter.

Relations 3.1.2 and 3.1.3 represent, of course, a constrained extremization problem; namely

$$\operatorname*{Max}_{\beta} J(\beta,\mathfrak{F},\beta,\mathfrak{F}),$$

with restriction 3.1.2 and optimizer α, and it seems useful to note that 3.1.2 and 3.1.3 explicitly relate objective function, restriction and optimizer via α and $\mathfrak{F}$. Thus $\mathfrak{F}$ determines the objective function (i.e., the particular mutual information functional in terms of which the β's are to compete), α is the optimizer, and α and $\mathfrak{F}$ together determine the restriction. Analogous remarks apply to 3.1.4 and 3.1.5, for which case the natural connection between solution and restriction also is evidenced in equation 4.2.14 of T. Berger [8], though only for a special version of 3.1.4 and 3.1.5, presumably with the special motivation of relating solution to restriction.

Lastly, it seems of profit to point out that restrictions 3.1.2 and 3.1.4 are in effect moment conditions; i.e., conditions that fix the

expectation of the "loss function"

$$\ln [dF_x/dF_\alpha(y)]. \tag{3.1.6}$$

Indeed, in the multivariate normal application of section 3, condition 3.1.2 specifies the nonnegativity of a certain linear function of the variances and covariances for β, and thus generalizes and replaces previously given conditions that specify the individual values of all of these variances and covariances (Kolmogorov [35], T. Berger [8]).

This "loss function" interpretation also underlies the illustrative computations of section 4. In a Bayesian sense, the loss functions given there are tied intrinsically to priors and likelihoods, via mutual information. By the same token, the priors are intrinsically tied to the loss functions and likelihoods, and thus are not without Bayesian interest.

We have couched the details of the ensuing discussion in terms of measures, rather than distribution functions, and section 2 contains some preliminaries to that effect. Sections 3 and 4 are devoted, respectively, to 3.1.2 and 3.1.3, and to 3.1.4 and 3.1.5.

2. Preliminaries

$M = (X,\mathcal{S})$ is a measurable space on which are given a pair of probability measures α and β, with $\alpha << \beta$ (recall that $\alpha << \beta$ if $\alpha(E) = 0$ for each $E \in \mathcal{S}$ for which $\beta(E) = 0$). Note that the phrase "almost all x" appearing below will refer to β (and hence as well to α). $N = (Y,\mathcal{T})$ is a second measurable space, and $\mathcal{F} = \{F_x\}$ is a collection of probability measures on N, indexed by the elements x of X. F_α and F_β are the probability measures on N given, for $\lambda = \alpha$ or β and $T \in \mathcal{T}$, by $F_\lambda(T) = \int_X F_x(T)d\lambda(x)$. (cf. Robbins [50] and Sethuraman [55] for discussion of such marginal probability measures and the corresponding joint probability measures.) Note that $\alpha << \beta$ implies $F_\alpha << F_\beta$, and we denote by $f_\beta^\alpha(y)$ the density $dF_\alpha/dF_\beta(y)$.

Analogous things are given for a family $\mathcal{G} = \{G_x\}$ on N, with the connective assumption that $F_x << G_\alpha$ for almost all x, and with the corresponding density $h_\alpha^x(y)$, which is $\mathcal{S} \times \mathcal{T}$-measurable.

In addition, assume (Regularity condition C) that:

Ca: $F_x << F_\alpha$ (and hence $F_x << F_\beta$) for almost all x;

Cb: the densities $f_\alpha^x(y)$ and $f_\beta^x(y)$ guaranteed under Ca for almost all x, are such that $F_x(y\ ;\ f_\lambda^x(y) \leq r)$, $\lambda = \alpha$ or β, is $\mathcal{S}$-measurable for all real r;

Cc: $f_\lambda^x(y)$, $\lambda = \alpha$ or β, is $\mathcal{S} \times \mathcal{T}$-measurable.

Analogous regularity conditions Da, Db, and Dc pertain to G. Finally assume (Regularity condition E) that

$$\int\int_{XY} \left| \ln \left[g_\alpha^x(y)/f_\alpha^x(y) \right] \right| g_\alpha^x(y)\, d\alpha(x)\, dG_\alpha(y) < +\infty. \tag{3.2.1}$$

Condition Ca is always satisfied when X is countable. There are, however, exceptions, as for example when M is the unit interval, α is uniform on M, and F_x concentrate its mass at the singleton $\{x\}$; i.e., for almost all x, $F_x(\{x\}) = 1$. Denote this class of probability measures by $\mathcal{F}^*$. For then, if Condition Ca held, we would have $F_\alpha(\{x\}) > 0$ for almost all x, so that F_α would have to be a measure on (0,1) assigning positive mass to almost all, i.e., uncountably many, singletons of (0,1). Note that this counter example, pinpointing the lack of separability of the class $\mathcal{F}^*$ with respect to the metric $d(F_{x_1},F_{x_2}) = \sup_{E \in \mathcal{T}} |F_{x_1}(E) - F_{x_2}(E)|$, $F_{x_1},F_{x_2} \in \mathcal{F}^*$, is consistent with the theorem of A. Berger [7]. (See also

p. 352 of Lehmann [38].)

Consider now the integrals

$$I(F_x,F_\alpha) = \int_Y \ln f_\alpha^x(y)dF_x(y) \tag{3.2.2}$$

and $I(F_x,F_\beta)$. These two integrals are well-defined and nonnegative (though possibly $+\infty$), in view of Theorem 4.1 of Bahadur [1]. They are further $\mathcal{A}$-measurable functions of x, as limits of measurable functions, in view of Cb. Hence, for $(\theta,\lambda) = (\alpha,\alpha)$, (β,β), (α,β), or (β,α), the functionals

$$J(\theta,\mathfrak{F},\lambda,\mathfrak{F}) = \int_X[\int_Y \ln f_\theta^x(y)dF_x(y)]d\lambda(x) \tag{3.2.3}$$

are well-defined and nonnegative (though possibly $+\infty$).

Analogous remarks pertain to $I(G_x,G_\alpha)$, $I(G_x,G_\beta)$ and $J(\theta,\mathfrak{B},\lambda,\mathfrak{B})$.

In the remaining part of the section, we list a theorem and two lemmas that will be of use in the later development.

Lemma 3.2.1: Let λ be a probability measure on $M = (X,\mathcal{A})$, and let $\mathcal{H} = \{H_x\}$ be a family of probability measures on $N = (Y,\mathcal{J})$, indexed by the elements x of X. Let ν be a σ-finite measure on $(Y,\mathcal{J})$ with $H_x << \nu$ almost all x, such that $dH_x/d\nu(y)$ is $\mathcal{A}\times\mathcal{J}$-measurable and let

H_λ be the measure on $(Y,\mathcal{J})$ defined by $H_\lambda(T) = \int_X H_x(T)d\lambda(x)$, $T\in\mathcal{J}$. Then $H_\lambda << \nu$, with

$$dH_\lambda/d\nu(y) = \int_X [dH_x/d\nu(y)]d\lambda(x), \quad \text{a.e.}[\nu]. \tag{3.2.4}$$

Proof: First note that, for $T\in\mathcal{J}$,

$$\int_T [\int_X dH_x/d\nu(y)d\lambda(x)]d\nu(y) \tag{3.2.5}$$

$$= \int_X [\int_T dH_x/d\nu(y)d\nu(y)]d\lambda(x) = \int_X H_x(T)d\lambda(x) = H_\lambda(T),$$

where the first equality is due to Tonelli's Theorem (cf. p. 270 of Royden [53]). The fact that $H_\lambda << \nu$ follows from 3.2.5 immediately. Now, in view of essential uniqueness of the Radon-Nykodym derivative, 3.2.4 follows.

Lemma 3.2.1, applied with $\nu = G_\alpha$, $H_x = F_x$, and $\lambda = \alpha$, shows that our earlier assumption $F_x << G_\alpha$ implies that also $F_\alpha << G_\alpha$; we denote the corresponding density by $h_\alpha(y)$. The next theorem, an "asymmetric version of Fubini's Theorem," will be of use in section 3.

Theorem 3.2.1 (Theorem 3 of Robbins [50]): Let M, N, $\mathcal{H}$ and H_λ be defined as in Lemma 3.2.1. Also let f(y) be a $\mathcal{J}$-measurable function on N with $f^+ = \max(f,0)$ and $f^- = \min(f,0)$. Then a necessary and

sufficient condition that

$$\int_Y f(y)dH_\lambda(y) = \int_X [\int_Y f(y)dH_x(y)]d\lambda(x) \tag{3.2.6}$$

is that at least one of the two quantities

$$\int_Y f^+(y)dH_\lambda(y) \tag{3.2.7}$$

and

$$\int_Y f^-(y)dH_\lambda(y) \tag{3.2.8}$$

be finite.

The following Kullback-Leibler information number-related inequality will be useful in section 4.

Lemma 3.2.2 (Lemma 1.1 of Csiszár [15]): Let $\phi(\cdot)$ be an arbitrary concave function defined in $(0,+\infty)$ with $\phi(0) = \lim_{u \to +0} \phi(u)$. Let $g(x)$ and $h(x)$ be two nonnegative measurable functions on a σ-finite measure space $(X,\mathcal{B},\xi)$; then $\int_S g(x)\phi[g(x)/h(x)]d\xi(x)$ is well-defined for all $S \in \mathcal{B}$ on which $g(x)$ and $h(x)$ are integrable, and for such an S,

$$\int_S g(x)\phi[g(x)/h(x)]d\xi(x) \geq \int_S g(x)d\xi(x) \cdot \phi[\int_S g(x)d\xi(x)/\int_S h(x)d\xi(x)] > -\infty.$$

3. The maximization problem

We now establish 3.1.3 under 3.1.2. Note first, in view of the discussion concerning 3.2.3, that $J(\alpha,\mathfrak{F},\beta,\mathfrak{F})$ and $J(\beta,\mathfrak{F},\beta,\mathfrak{F})$ both are

well-defined, though possibly $+\infty$. Moreover the former is in fact finite, because of 3.1.2. Also, in view of the discussion concerning 3.2.2, $I(F_x,F_\alpha)$ and $I(F_x,F_\beta)$ both are well-defined for almost all x, and $I(F_x,F_\alpha)$ must be finite for almost all x, since otherwise $J(\alpha,\mathfrak{F},\beta,\mathfrak{F})$ could not be finite. These remarks validate the first two equalities of 3.3.1 below, and thus we write

$$
\begin{aligned}
& J(\beta,\mathfrak{F},\beta,\mathfrak{F}) - J(\alpha,\mathfrak{F},\beta,\mathfrak{F}) \\
&= \int_X [I(F_x,F_\beta) - I(F_x,F_\alpha)]d\beta(x) \\
&= \int_X [\int_Y \ln [f_\beta^x(y)/f_\alpha^x(y)]dF_x(y)]d\beta(x) \\
&= \int_X [\int_Y \ln [f_\beta^\alpha(y)]dF_x(y)]d\beta(x) \qquad (3.3.1) \\
&= \int_Y \ln [f_\beta^\alpha(y)]dF_\beta(y) \\
&\leq 0,
\end{aligned}
$$

where the third equality is due to the chain rule, and the last equality is due to Theorem 3.2.1 with expression 3.2.8 in fact finite, since $J(\alpha,\mathfrak{F},\beta,\mathfrak{F})$ is finite. The inequality is due to Theorem 4.1 of Bahadur [1].

One practical off-shoot of the fact that 3.1.3 holds under 3.1.2 occurs when $(\alpha,\mathfrak{F})$ is $(m + n)$-normal, with α nonsingular m-normal: $\alpha = N(0,\Sigma)$, and F_x nonsingular n-normal, with linear regressions and constant variance-covariance matrix: $F_x = N(Bx, \Theta)$. Then the quantity 3.1.6 becomes;

$$\begin{aligned}
&\ln [dF_x/dF_\alpha(y)] \\
&= \tfrac{1}{2} \ln [|\Sigma||\ \Theta + B\Sigma B'|/|\Sigma_F|] \\
&-\tfrac{1}{2}[\binom{x}{y}'\Sigma_F^{-1}\binom{x}{y} - x'\Sigma^{-1}x - y'(\ \Theta + B\Sigma B')^{-1}y\} \\
&= K_0 - \tfrac{1}{2}\{tr[\binom{x}{y}\binom{x}{y}'\Sigma_F^{-1}] - tr[xx'\Sigma^{-1}] - tr[yy'(\ \Theta + B\Sigma B')^{-1}]\},
\end{aligned}$$

where F and Σ_F denote the joint $(m + n)$-normal probability measure and the corresponding variance-covariance matrix, and

$$K_0 \equiv \tfrac{1}{2} \ln [|\Sigma||\ \Theta + B\Sigma B'|/|\Sigma_F|].$$

Now, since $J(\alpha,\mathfrak{F},\alpha,\mathfrak{F}) = E_{\alpha,\mathfrak{F}}\{\ln [dF_x/dF_\alpha(y)]\} = K_0$, letting μ and V denote the first and second moments of β, the relation 3.1.2 takes the form: $J(\alpha,\mathfrak{F},\beta,\mathfrak{F}) = E_{\beta,\mathfrak{F}}\{\ln [dF_x/dF_\alpha(y)]\} = K_0$, namely,

$$tr\left[\begin{pmatrix} V & VB' \\ BV & \Theta + BVB' \end{pmatrix}\begin{pmatrix} \Sigma & \Sigma B' \\ B\Sigma & \Theta + B\Sigma B' \end{pmatrix}^{-1}\right] \qquad (3.3.2a)$$

$$- tr[V\Sigma^{-1}] - tr[(\ \Theta + BVB')(\ \Theta + B\Sigma B')^{-1}] \qquad (3.3.2b)$$

$$= \mu'[B'(\Theta + B\Sigma B')^{-1}B]\mu. \qquad (3.3.2c)$$

Since J does not depend on μ, and since the quadratic form of 3.3.2c is positive semi-definite, μ may be thought of as acting as a slack for V, and 3.1.2 reduces to the inequality 3.3.2a + 3.3.2b $\geq$ 0, which reduces in turn to $V \leq \Sigma$ when m = n = 1. Thus $N(0,\Sigma)$ is an m-dimensional distribution maximizing $J(\beta,\mathfrak{F},\beta,\mathfrak{F})$ under this single inequality. This fact generalizes the assertion, on page 106 of Kolmogorov [35] that $N(0,\Sigma)$ maximizes $J(\beta,\mathfrak{F},\beta,\mathfrak{F})$ under the [m(m + 1)/2] moment equalities $V = \Sigma$; this last assertion itself being the m-dimensional generalization of Theorem 4.3.4 of T. Berger [8].

4. The minimization problem

We now establish 3.1.5 under 3.1.4. Again, the expressions $J(\alpha,\mathfrak{F},\alpha,\mathscr{E})$ and $J(\alpha,\mathscr{L},\alpha,\mathscr{E})$ of 3.1.4 and 3.1.5 are well-defined, in view of the discussion preceding 3.2.3. Moreover both may now be taken as finite; the first because of 3.1.4, and the second because, in view of 3.1.5, no generality is lost thereby. Analogously to section 3, these remarks validate the first two equalities of 3.4.1 below, and we write:

$$J(\alpha, \mathcal{G}, \alpha, \mathcal{G}) - J(\alpha, \mathcal{F}, \alpha, \mathcal{G})$$

$$= \int_X [\int_Y \ln g_\alpha^x(y) dG_x(y) - \int_Y \ln f_\alpha^x(y) dG_x(y)] d\alpha(x)$$

$$= \int_X [\int_Y \ln [g_\alpha^x(y)/f_\alpha^x(y)] dG_x(y)] d\alpha(x)$$

$$= \int_X [\int_Y \{\ln [g_\alpha^x(y)/f_\alpha^x(y)]\} g_\alpha^x(y) dG_\alpha(y)] d\alpha(x) \qquad (3.4.1)$$

$$= \int_Y [\int_X \{\ln [g_\alpha^x(y)/f_\alpha^x(y)]\} g_\alpha^x(y) d\alpha(x)] dG_\alpha(y)$$

$$= \int_Y [\int_X \{\ln [g_\alpha^x(y) \cdot h_\alpha(y)/h_\alpha^x(y) \cdot 1]\} g_\alpha^x(y) d\alpha(x)] dG_\alpha(y)$$

$$= \int_Y [\int_X \{\ln [g_\alpha^x(y)/h_\alpha^x(y)] - \ln [\int g_\alpha^s(y) d\alpha(s)/\int h_\alpha^s(y) d\alpha(s)]\}$$

$$g_\alpha^x(y) d\alpha(x)] dG_\alpha(y)$$

$$\geq 0.$$

Here the third equality follows from the definition of $g_\alpha^x(y)$, the fourth equality follows from Condition E and Fubini's Theorem, the fifth equality comes from the chain rule, the last equality is due to Lemma 3.2.1, and the inequality comes from applying Lemma 3.2.2 to the integrand.

Our applications make use of the fact that the restriction 3.1.4 clearly is equivalent to the restriction:

$$\int_X [\int_Y L(x,y) dG_x(y)] d\alpha(x) = \int_X [\int_Y L(x,y) dF_x(y)] d\alpha(x), \qquad (3.1.4^*)$$

where

$$L(x,y) = C(\ln f_\alpha^x(y) - t(x)),\ C \neq 0,$$

with $t(x)$ an arbitrary α-integrable function of x. When $\ln f_\alpha^x(x)$ is α-integrable and

$$t(x) = \ln f_\alpha^x(x) \tag{3.4.2}$$

one has $L(x,x) = 0$, and those cases are of special interest for which, in addition, the above choice of $t(x)$ also leads, with suitably chosen sign for C, to

$$L(x,y) \geq 0 \tag{3.4.3}$$

Our three examples, all with x and y scalar, satisfy 3.4.2 and 3.4.3 and have the additional feature that $L(x,y)$, for fixed x, is monotone in $|y - x|$ on either side of x, so that the restriction 3.1.4* takes on the appearance of a restriction on expected loss.

Our first example appears, essentially, in Section 4.3.3 of T. Berger [8].

Example A.

$$\alpha = N(0,1)$$

$$F_x = N(bx, b(1-b)),\ 0 < b < 1$$

$$\ln f_\alpha^x(y) = -\tfrac{1}{2} \ln (1-b) - \frac{(y-x)^2}{2(1-b)} - \frac{x^2}{2}$$

$$t(x) = -\tfrac{1}{2} \ln (1-b) - \frac{x^2}{2}$$

$$C = -2(1-b)$$

$$L(x,y) = (y-x)^2$$

Example B.

α is inverse binomial:

$$\alpha(x) = \binom{x + a-1}{a-1} p^x (1-p)^a;\ x:\ 0,1,2,\ldots$$

F_x is gamma:

$$f_x(y) = [\Gamma(x+a) p^{x+a}]^{-1} y^{x+a-1} e^{-y/p};\ y \geq 0$$

$$\ln f_\alpha^x(y) = \ln [\Gamma(a)/(1-p)^a] - \ln [\Gamma(x+a) \cdot p^x] - [y - x \ln y]$$

$$\equiv \theta(x) - [y - x \ln y]$$

$$t(x) = \theta(x) - [x - x \ln x]$$

$$C = -1$$

$$L(x,y) = (y-x) - x \ln (y/x)$$

Example C.

α is the binary inverse hypergeometric:

$$\alpha(x) = (a+x-1)\,!\,(b-x)\,!\,[(a-1)\,!\,(b-1)\,!\,(a+b)]^{-1};\ x = 0 \text{ or } 1$$

F_x is beta:

$$f_x(y) = B(x+a,\ b-x+1)^{-1} y^{x+a-1}(1-y)^{b-x};\ 0 \leq y \leq 1$$

$$\ln f_\alpha^x(y) = \ln [(a-1)!\,(b-1)!(a+b)/\{(a+x-1)\,!\,(b-1)\,!\}]$$

$$+ x \ln y + (1-x) \ln (1-y)$$

$$\equiv \theta(x) + x \ln y + (1-x) \ln (1-y)$$

$$t(x) = \theta(x) + x \ln x + (1-x) \ln (1-x)$$

$$= \theta(x)$$

$$C = -1$$

$$L(x,y) = -[x \ln y + (1-x) \ln (1-y)]$$

IV. MARKOV PROCESSES AS SPECIFIC FREE ENERGY MINIMIZERS

Based on the work of Lanford and Ruelle [37], Spitzer [56] considered the following problem of characterizing a Markov chain: Let $(\Omega,\mathcal{A},\mu)$ denote a stochastic process where $\Omega = S^{\mathbb{Z}}$, S is finite, and $\mathcal{A}$ is the σ-algebra of subsets of Ω generated by the cylinder sets. Also let $\mathcal{E}$ denote the class of stochastic processes that are stationary. Define the specific entropy associated with μ by

$$s(\mu) \equiv \lim_n n^{-1} \sum_x - \mu_n(x) \ln \mu_n(x), \quad \mu \in \mathcal{E}, \tag{4.1}$$

where μ_n is the restriction of μ to the first n-coordinates. Also define the specific energy associated with μ and a nearest neighbor potential $U(x,y)$ by

$$e_u(\mu) \equiv \lim_n n^{-1} \sum_x \mu_n(x) \sum_1^{n-1} U(x_i, x_{i+1}). \tag{4.2}$$

Then it is shown in [56] that a certain Markov chain associated with $U(x,y)$ is the stochastic process that minimizes the specific free energy

$$f_u(\mu) = e_u(\mu) - s(\mu) \text{ over } \mathcal{E}. \tag{4.3}$$

Motivated by this result (which is in fact a one-dimensional specialization of [37]), we are interested in establishing an analogue of the easy noncharacterizing part of the theorem: The analogous stationary Markov process associated with a similarly restricted $U(x,y)$ minimizes the specific free energy, among the class $\mathfrak{F}$ of all (non-necessarily-stationary) processes μ possessing n-dimensional marginal densities f_n on the products $(S^n, \mathcal{B}^n, \nu^n)$ of a finite measure space $(S, \mathcal{B}, \nu)$.

First of all, define, for $\mu \in \mathfrak{F}$,

$$e_u^n(\mu) \equiv \int \{ \sum_1^{n-1} U(x_i, x_{i+1}) + \ln (f_n(x^n)) \} f_n(x^n) d\nu^n, \tag{4.4}$$

where $x^n \equiv (x_i, \ldots, x_n)$ and $U(x,y)$ is ν^2-integrable. Also consider a μ^* with n-dimensional marginal density of form

$$f_n^* = C_n \exp\{- \sum_1^{n-1} U(x_i, x_{i+1})\}. \tag{4.5}$$

Then, for each fixed n,

$$e_u^n(\mu) \geq \ln C_n, \ \forall \mu \in \mathfrak{F} \tag{4.6}$$

and μ^* satisfies 4.6 with equality. In fact, to see 4.6, one only needs to observe that $\int \ln (f_n / f_n^*) f_n d\nu^n \geq 0$, since the LHS is a Kullback-Leibler

information number. That

$$e_u^n(\mu^*) = \ln C_n \tag{4.7}$$

is a matter of substitution.

Next, it remains to compare $e_u^n(\mu^*)$ to $e_u^n(\mu^{**})$, where μ^{**} is an appropriate stationary Markov process. To this end assume that the kernel

$$Q(x,y) \equiv \exp\{-U(x,y)\}$$

on S^2 possesses left and right eigenfunctions $\phi(x)$ and $\psi(y)$, corresponding to a positive eigenvalue λ, that satisfy

$$\phi\psi \in \mathfrak{L}^2(S,\mathcal{B},\nu), \tag{4.8}$$

$$\phi(x),\ \psi(x) \geq \tau > 0 \text{ on } S, \tag{4.9}$$

and are normalized such that

$$\int \phi(x)\psi(x)d\nu = 1,$$

and let μ^{**} be the Markov process with initial probability density $\phi(x)\psi(x)$ and transition probability density kernel $Q(x,y)\psi(y)/\lambda\psi(x)$. Then the n-dimensional marginal density for μ^{**} is

$$f_n^{**} = \phi(x_1)\psi(x_n)\exp\{-\sum_1^{n-1} U(x_i,x_{i+1})\}/\lambda^{n-1} \tag{4.10}$$

and

$$e_u^n(\mu^*) - e_u^n(\mu^{**})$$

$$= \ln C_n + (n-1) \ln \lambda - \int\{\ln \phi(x_1) + \ln \psi(x_n)\} f_n^{**}(x^n) d\nu^n. \quad (4.11)$$

But both f_n^* and f_n^{**} integrate to 1 on S^n, so that using 4.9,

$$\ln C_n + (n-1) \ln \lambda \geq \ln \tau^2 \quad (4.12)$$

and, $\int\{\ln \phi(x_i) + \ln \psi(x_n)\} f_n^{**}(x^n) d\nu^n$

$$= \int\{\ln \phi(x_1)\} \phi(x_1)\psi(x_1) d\nu + \int\{\ln \psi(x_n)\} \phi(x_n)\psi(x_n) d\nu$$

$$= \int\{\ln(\phi(t)\psi(t))\} \phi(t)\psi(t) d\nu \leq \ln \int\{\phi(t)\psi(t)\}^2 d\nu \equiv I < +\infty, \quad (4.13)$$

where the last strict inequality is due to 4.8. Hence applying 4.12 and 4.13 to 4.11 yields

$$e_u^n(\mu^*) - e_u^n(\mu^{**}) \geq \ln \tau^2 - I,$$

which, together with 4.7 and 4.8, yields, for $\mu \in \mathfrak{F}$,

$$e_u^n(\mu) \geq e_u^n(\mu^*) \geq e_u^n(\mu^{**}) + \ln \tau^2 - I,$$

so that, for $\mu \in \mathfrak{F}$,

$$\underline{\lim}_n n^{-1} e_u^n(\mu) \geq \lim_n n^{-1} e_u^n(\mu^{**}) = -\ln \lambda, \quad (4.14)$$

where the equality is due to 4.7, 4.10, and 4.13.

Note that the first step of the argument, featuring μ^*, can either be phrased explicitly, as in Chapter I, in terms of our primal-Lagrangian approach, or implicitly, as above. A third approach, which almost eliminates the second step, proceeds by showing directly that

$$\underline{\lim}_n\, n^{-1} e_u^n(\mu) \geq -\ln \lambda, \quad \forall \mu \in \mathfrak{F}. \tag{4.15}$$

To this end, note first that, for $s > 0$, $t > 0$,

$$s \ln s - t \ln t \geq (s - t)(\ln t + 1), \tag{4.16}$$

which follows immediately from $1 - t/s \leq \ln (s/t)$. Then,

$$\begin{aligned}
& n^{-1} e_u^n(\mu) + \ln \lambda \\
&= n^{-1}\{\int [\textstyle\sum_1^{n-1} U(x_i, x_{i+1}) + \ln (f_n(x^n))] f_n(x^n) d\nu^n - \ln C_n\} \\
&\qquad + n^{-1} \ln C_n + \ln \lambda \\
&= n^{-1}\{\int [\Sigma U + \ln f_n] f_n d\nu^n - \int [\Sigma U + \ln f_n^*] f_n^* d\nu^n\} + n^{-1} \ln C_n + \ln \lambda \\
&\geq n^{-1}\{\int (f_n - f_n^*)(\ln f_n^* + 1) d\nu^n + \int (f_n - f_n^*)(\Sigma U) d\nu^n\} + n^{-1} \ln C_n + \ln \lambda \\
&= n^{-1} \ln C_n + \ln \lambda \geq n^{-1} \ln (\tau^2 \lambda)
\end{aligned} \tag{4.17}$$

where the first equality comes from adding and subtracting $n^{-1} \ln C_n$, the second and third equalities come from 4.5, the first inequality

comes from 4.16, with $f_n = s$ and $f_n^* = t$, and the last inequality comes from 4.12. Finally, 4.15 follows from 4.17.

V. A GENERALIZED TCHEBYCHEFF PROBLEM

1. Introduction

C. R. Mischke [42] has posed the following problem: Imagine an object operating under stress. Suppose that the strength distribution of the object is not known completely, whereas the stress to which it is subject, acting negatively to the strength in a linear fashion, is assumed to have a known distribution function. We want to find the maximum probability of reliable performance of the object, i.e., the maximum probability that strength exceeds stress. Casting the problem in general terms, let X and Y be a pair of independent random variables. Assume that the distribution function of Y, denoted by $G(\cdot)$, is completely specified, while the distribution function of X, denoted by $F(\cdot)$, is unknown except for a set of several moments of X. Compute the upper bound for $\Pr\{Y<X\}$.

Note that this problem is a slight variation of the classical problem underlying the Tchebycheff inequality. In fact, if a symmetric set characteristic function replaces $G(\cdot)$ appearing in the integral $\int G(t)dF(t) = \Pr\{Y<X\}$, and if maximization is subject to given values of

the first two moments on X, then the problem is reduced to the original question of Tchebycheff. Indeed, problems of both types, i.e., the classical Tchebycheff problem and the above strength-stress problem, are special cases of what in Karlin and Studden [26] are called "generalized Tchebycheff problems". Related problems, but not in the spirit of the Tchebycheff inequality, of finding the minimum variance unbiased estimate of and the confidence interval for $Pr\{Y<X\}$ based on a random sample from X, have been considered in the literature, for example, by Church and Harris [12], with parametric conditions on X and Y, and by Govindarajulu [20] for the nonparametric case.

The fact that problems of Tchebycheff type can be solved effectively in the framework of the duality theory of mathematical programming has been documented in Isii [25], Karlin and Studden [26], Whittle [62], and Pyne [47] among others. (See also Kingman [34] and Kemperman [27] for similar treatments, but without explicit recourse to mathematical programming framework.) To couch the discussion in this format, first define a class of functions $\mathfrak{F} = \{F | F$ a nonnegative, nondecreasing, and right

continuous function of bounded variation, defined on $T \subset E_1$}. For a set $h(t) \equiv \{h_i(t)\}_{i=0}^{n}$ of piecewise continuous functions on T, let $CH[h(T)] \subset E_{n+1}$ denote the convex hull generated by { $h(t)$, $t \in T$}. Assume that $b = (b_0, b_1, \ldots, b_n)' \in CH[h(T)]$ to avoid trivial inconsistency. Now consider the program;

$$P: \quad \sup \int G(t)dF(t) \tag{5.1.1}$$

$$\text{s.t.} \quad \int h_i(t)dF(t) = b_i, \quad 0 \leq i \leq n, \tag{5.1.2}$$

$$\text{and} \quad F \in \mathfrak{F}. \tag{5.1.3}$$

Here we set $h_0(t) \equiv 1 \equiv b_0$ to normalize F.

Associated with P, using the dual cone structure corresponding to 5.1.2 and 5.1.3, (cf. Sposito [57] for the finite case), one may write down a program formally dual to P;

$$D: \quad \inf \lambda \cdot b \tag{5.1.4}$$

$$\text{s.t.} \quad \lambda \cdot h(t) \geq G(t), \quad \forall t \in T, \tag{5.1.5}$$

$$\text{and} \quad \lambda = (\lambda_0, \lambda_1, \ldots, \lambda_n) \in E_{n+1}. \tag{5.1.6}$$

In passing, note that D is consistent since, for example, we can choose $\lambda_0 = 1$ and $\lambda_i = 0$, $1 \leq i \leq n$, satisfying 5.1.5 and 5.1.6.

Two procedures are used in this chapter to analyze the pair (P,D). The first of these, in the spirit of the previous chapters, establishes the weak duality for the pair (P,D), and then proceeds to verify, with a pair of candidate solutions for (P,D), that the two objective functions of P and D in fact attain a common value. As a result, we can obtain not only a sharp bound for $\int G(t)dF(t)$ but also an extremal random variable that achieves this bound. Section 2 contains a further discussion of this procedure, including applications to two illustrative problems.

For the second procedure, which will be pursued in section 3, it seems to be useful to incorporate the normalization $\int dF(t) = 1$ into the underlying space $\mathfrak{F}$. Then the modified program is;

$$P_1: \quad \sup \int G(t)dF(t)$$

$$\text{s.t.} \quad \int h_i(t)dF(t) = b_i, \quad 1 \leq i \leq n,$$

$$\text{and} \quad \mathfrak{F}_0 = \{F \mid F \in \mathfrak{F} \text{ and } \int dF(t) = 1\}$$

Now, define $\mathfrak{F}_d = \{F \mid F \text{ a discrete c.d.f. with at most } n+1 \text{ jumps}\}$, a subset of $\mathfrak{F}_0$. Then, it has been shown that the (optimal) value of the program P_1 remains unaltered, when $\mathfrak{F}_0$ is replaced by $\mathfrak{F}_d$ (Richter [48],

Rogosinski [52], and Mulholland and Rogers [44]). In view of this, we will reformulate P_1 in terms of $\mathfrak{F}_d$, and denote this reformulated program by P_2. The reason we prefer the nonconal convex sets $\mathfrak{F}_0$ of P_1 and $\mathfrak{F}_d$ of P_2 for the second procedure can be explained, at least partially, by the fact that the members of $\mathfrak{F}_d$ serve as generators of a certain convex hull. Banking on the characterization of consistency of P_2 in terms of the set $C_o \equiv CH[\Gamma_o]$, where $\Gamma_o = \{x = (x_1, \ldots, x_n) | x_i = h_i(t),\ 1 \leq i \leq n$ for some $t \epsilon T\}$, as provided by Lemma 2 of Kemperman [27] (which follows from the theorem of Carathéodory), we write down the third formulation P_3, equivalent to P_2, in the spirit of Van Slyke and Wets [59] and Wets [61], as follows.

$$P_3: \sup_{x \epsilon C \cap \mathfrak{L}_b} x_{n+1}$$

where

$$C = CH[\Gamma],\ \Gamma = \{x = (x_1, \ldots, x_{n+1}) | x_i = h_i(t),\ 1 \leq i \leq n,$$
$$x_{n+1} = G(t), \text{ for some } t \epsilon T\},$$

$$\text{and } \mathfrak{L}_b = \{x = (x_1, \ldots, x_{n+1}) | x_i = b_i,\ 1 \leq i \leq n,\ x_{n+1} \epsilon E_1\}.$$

A detailed demonstration of the equivalence of P_2 and P_3 may be found in Pyne [47]. Section 3 begins with an analysis of the program P_3, and

treats the same concrete problems discussed in section 2. In section 4, a comparison is made between the two procedures of sections 2 and 3.

2. Weak duality approach

It is elementary to check that $\int G(t)dF^{o}(t) \leq \lambda^{o} \cdot b$ for a feasible pair (F^{o}, λ^{o}) of solutions for the pair (P,D), i.e., that (P,D) is weakly dual. As a consequence of this, we may conclude that (F^{*},λ^{*}) is an optimal pair of solutions for the program pair (P,D) if

$$\int\{\lambda^{*} \cdot h(t) - G(t)\}dF^{*}(t) = 0 . \tag{5.2.1}$$

Now, from the constraint 5.1.5 of D, define a "contact" set, with λ^{*} in 5.2.1,

$$\mathfrak{J} \equiv \{t\epsilon T \,|\, \lambda^{*} \cdot h(t) - G(t) = 0\}. \tag{5.2.2}$$

Assuming that there is a λ^{*} such that $\mathfrak{J}$ is nonempty, if we can construct an F^{*} whose mass is concentrated on the set $\mathfrak{J}$, then (F^{*},λ^{*}) is an optimal pair in view of 5.2.1. Of course it is not necessarily true that every point in $\mathfrak{J}$ should have positive mass. Note that, due to the general polynomial structure of the function $\lambda^{*} \cdot h(t) - G(t)$, a crude upper bound for the cardinality of $\mathfrak{J}$ is n + 2, which, in turn, gives an upper bound

for the number of points in the spectrum of the extremal distribution F^*.

Note that, in the spirit of the previous chapters, the discussion so far is free of regularity condition. A sufficient condition appearing frequently for the existence of λ^*, and hence the equality of the optimal values for P and D, (sometimes called "strong duality"), is that the RHS vector b in P is contained in the interior of the moment space spanned by $\{h_i(t)\}_{i=0}^{n}$, and this fact is intrinsically tied to the linear independence of $\{h_i(t)\}$ with respect to the domain T. Although this type of condition is useful when one wants to check whether there is a solution to a given problem or not, it does little toward providing a scheme for reaching an optimal solution.

Ex. 1 We wish to find the maximum reliability when the stress distribution is a Laplacian and the strength distribution is known up to the first two moments. Namely, we consider the program

$$
\begin{aligned}
P: \quad & \sup \int G(t)dF(t) \\
& \text{s.t.} \int dF(t) = 1 \\
& \qquad \int t\,dF(t) = b_1 \\
& \qquad \int t^2 dF(t) = b_2 \\
& \text{and } F \in \mathfrak{F}.
\end{aligned}
$$

For computational simplicity, let $b_1 = 0$ and let $G(t) = \frac{1}{2}e^{-|t|}$ for $t < 0$,

$1 - \frac{1}{2}e^{-t}$ for $t \geq 0$.

A formal dual program to P is

$$\begin{aligned} D: \quad & \inf_{\lambda} \quad \lambda_0 + \lambda_1 b_1 + \lambda_2 b_2 \\ & \text{s.t.} \quad \lambda_0 + \lambda_1 t + \lambda_2 t^2 \geq G(t), \ \forall t \in E_1, \\ & \text{and} \quad \lambda = (\lambda_0, \lambda_1, \lambda_2) \in E_3. \end{aligned}$$

Now, analogously to 5.2.2, denote $\mathfrak{T} = \{t \mid \lambda_0^* + \lambda_1^* t + \lambda_2^* t^2 = G(t)\}$, and let t_1 and t_2, $t_1 < 0 < t_2$, be in $\mathfrak{T}$. The fact that an extremal distribution should concentrate its mass on $\mathfrak{T}$, and the feasibility for P, yield

$$t_1 p + t_2(1 - p) = 0, \tag{a}$$

and

$$t_1^2 p + t_2^2(1 - p) = b_2, \tag{b}$$

where $p \in (0,1)$.

Moreover, $t_1 \in \mathfrak{T}$ and $t_2 \in \mathfrak{T}$ trivially yield,

$$\lambda_0 + \lambda_1 t_1 + \lambda_2 t_1^2 = \tfrac{1}{2}e^{-|t_1|}, \tag{c}$$

and

$$\lambda_0 + \lambda_1 t_2 + \lambda_2 t_2^2 = 1 - \tfrac{1}{2}e^{-t_2}. \tag{d}$$

The equations for the first derivatives obtained from (c) and (d) are,

$$\lambda_1 + 2\lambda_2 t_1 = \tfrac{1}{2}e^{-|t_1|}, \tag{c'}$$

and

$$\lambda_1 + 2\lambda_2 t_2 = \tfrac{1}{2}e^{-t_2}. \tag{d'}$$

Finally, the equation for optimality, analogous to 5.2.1, yields,

$$\tfrac{1}{2}e^{-|t_1|} \cdot p + (1 - \tfrac{1}{2}e^{-t_2}) \cdot (1 - p) = \lambda_0 + \lambda_2 b_2 \ . \qquad \text{(e)}$$

We wish to solve the above equations (a)-(e) for (p,t_1,t_2) and $(\lambda_0,\lambda_1,\lambda_2)$, and, as sketched below, the computation is manageable.

Sketch of a solution

(i) (a) + (b) $\rightarrow$ $p = -t_2/(t_1 - t_2)$ (1)

$$t_1 = -b_2/t_2, \quad t_2 = -b_2/t_1 \qquad (2)$$

$$(1) + (2) \rightarrow p = t_2^2/(b_2 + t_2^2) \qquad (3)$$

(ii) (c) + (c′) + (2) $\rightarrow$ $\lambda_0 = \tfrac{1}{2}e^{-b_2/t_2}(1 + b_2/t_2) + \lambda_2 \cdot b_2^2/t_2^2$ (4)

(d) + (d′) $\rightarrow$ $\lambda_0 = 1 - \tfrac{1}{2}e^{-t_2}(1 + t_2) + \lambda_2 t_2^2$ (5)

(iii) (e) + (4) + (2) + (3) $\rightarrow$

$$\tfrac{1}{2}e^{-b_2/t_2} \cdot t_2^2/(b_2 + t_2^2) + (1 - \tfrac{1}{2}e^{-t_2}) \cdot b_2/(b_2 + t_2^2)$$

$$= \tfrac{1}{2}e^{-b_2/t_2} \cdot (t_2 + b_2)/t_2 + \lambda_2(b_2^2/t_2^2 + b_2) \qquad (6)$$

(e) + (5) + (3) $\rightarrow$

$$\tfrac{1}{2}e^{-b_2/t_2} \cdot t_2^2/(b_2 + t_2^2) + (1 - \tfrac{1}{2}e^{-t_2}) \cdot b_2/(b_2 + t_2^2)$$

$$= 1 - \tfrac{1}{2}e^{-t_2}(1 + t_2) + \lambda_2(t_2^2 + b_2) \qquad (7)$$

(iv) (6) + (7) $\rightarrow$ $e^{-b_2/t_2} + e^{-t_2} = 4t_2/(t_2^2 + 2t_2 + b_2)$ (8)

Now, once (8) is solved for t_2, we can find p, t_2, λ_0, λ_1, and λ_2 using the above equations (a)-(e) and (1)-(7).

Ex. 2 We wish to find maximum reliability when the stress distribution belongs to the class

$$\mathcal{G} = \{G(\cdot);\ \text{strictly convex on } (-\infty,0) \text{ and strictly concave on } [0,\infty)\}, \tag{5.2.3}$$

and the strength distribution is known up to the first moment and the absolute moment. (Note here that, as will be seen later, this example illustrates the case for which the procedure via weak duality is less recommendable, because there does not exist an extremal random variable with which we can verify 5.2.1, except when the RHS specifies a boundary point of the appropriate moment space.) Namely we consider the program

$$\text{P:}\quad \sup \int G(t)dF(t)$$

$$\text{s.t.}\quad \int dF(t) = 1$$

$$\int t\,dF(t) = b_1 \tag{5.2.4}$$

$$\int |t|\,dF(t) = b_2 \tag{5.2.5}$$

$$\text{and}\quad F \in \mathfrak{F}.$$

A formal dual program to P is;

$$\text{D:} \quad \inf_{\lambda} \ \lambda_0 + \lambda_1 b_1 + \lambda_2 b_2$$

$$\text{s.t.} \quad \lambda_0 + \lambda_1 t + \lambda_2 |t| \geq G(t), \ \forall t \in E_1,$$

$$\text{and} \quad \lambda = (\lambda_0, \lambda_1, \lambda_2) \in E_3.$$

Analogously to 5.2.2, let

$$\mathfrak{I} = \{t | \lambda_0^* + \lambda_1^* t + \lambda_2^* |t| = G(t)\}. \tag{5.2.6}$$

Then, in view of strict concavity of $G(t)$ for $t \geq 0$, $\mathfrak{I}$ is singleton with a nonnegative t^*, which, in turn, implies that we have to confine ourselves to the set of degenerate (at t^*) c.d.f.'s. This fact is not necessarily helpful in constructing a candidate solution F^* to verify the equality analogous to 5.2.1, since in general such an F^* is inconsistent with the specified values (b_1, b_2). Only when $b_1 = b_2 > 0$, i.e., (b_1, b_2) is a boundary point in E_2^+ of the moment space M spanned by $\{t, |t|\}$, can we locate a consistent F^* using 5.2.6, and in this case one of the two restrictions 5.2.4 and 5.2.5 is redundant. So in Ex. 2, the specification 5.2.6 is useful for a boundary point rather than an interior point of M. But one may argue that the problem with a boundary point

specification is not a programming problem since there is only one feasible solution for P, and this is trivially optimal.

3. Moment space approach

Exploiting the geometric overtone of the program P_3 formulated in section 1, the value of the program P_3 may be computed by associating this value with a supporting hyperplane of the convex set C in E_{n+1} at (b,x^*_{n+1}), an upper boundary point of C. But since the set $C \equiv CH[\Gamma]$ is given only in terms of Γ in P_3, we pursue the appropriate supporting hyperplane of C rather indirectly in the following sense; based on the fact that a set in E_m and its convex hull share the same supporting hyperplane, we take the view of identifying the supporting hyperplane of C in question as the limit of the sequence of hyperplanes in E_{n+1} which touch or cut the set Γ, which is given explicitly in P_3. In particular, where $\Gamma \subset E_3$ as in Ex. 1 and Ex. 2 of section 2, the finding of an 'optimal' hyperplane for Γ amounts to selecting the 'best' triple of points from the collection of triples of distinct points in Γ satisfying certain conditions. In what follows, we reduce the search to a collection of pairs of distinct points

in Γ. We treat the two examples of section 2 in reverse order here for reasons to become clear later on.

Ex. a (Ex. 2 of section 2)

$$P_3: \sup_{x \in C \cap \mathcal{L}_b} x_3$$

where $C = CH[\Gamma]$,

$\mathcal{L}_b = \{x = (x_1,x_2,x_3) | x_i = b_i,\ i = 1,2,\ x_3 \in E_1\}$ with

$b = (b_1,b_2)$, and $G(\cdot) \in \mathcal{G}$ defined in 5.2.3.

Here, $\Gamma = \{(x_1,x_2,x_3) | x_1 = t,\ x_2 = |t|,\ x_3 \leq G(t);\ \text{some } t \in E_1\}$.

Define the following subsets of Γ;

$$\Gamma_{u-} = \{(x_1,x_2,x_3) | x_1 = t,\ x_2 = |t|,\ x_3 = G(t);\ \text{some } t < 0\},$$

$$\Gamma_{u+} = \{(x_1,x_2,x_3) | x_1 = t,\ x_2 = |t|,\ x_3 = G(t);\ \text{some } t \geq 0\}$$

$$\Gamma_{\ell-} = \{(x_1,x_2,0) | x_1 = t,\ x_2 = |t|;\ \text{some } t < 0\},$$

$$\Gamma_{\ell+} = \{(x_1,x_2,0) | x_1 = t,\ x_2 = |t|;\ \text{some } t \geq 0\},$$

and denote $\Gamma_u = \Gamma_{u-} \cup \Gamma_{u+}$, $\Gamma_\ell = \Gamma_{\ell-} \cup \Gamma_{\ell+}$.

Now let Z represent a triple $\{z_1,z_2,z_3\}$ of distinct points in Γ_u, with $z_i = (x_{1i},x_{2i},x_{3i})$, $1 \leq i \leq 3$. For any point $z \in E_3$, we consider the projection L of z into the plane $x_3 = 0$, denoted as $z_\ell \equiv L(z)$.

Similarly the projection U of z into the line $x_1 = x_2 = 0$, denoted as $z_u \equiv U(z)$. Also used are shortcut notations such as

$Z_\ell = L(Z)$, the projection L of a triple Z,

$Y_\ell = L(Y)$, the projection L of a pair Y, and so on.

Now, suppose that a point $b \in CH[\Gamma_\ell]$ is given, and state the condition Q on Z;

Condition Q: $b \in CH[Z_\ell]$,

and define these Z-sets;

S: $\{Z \mid Z \text{ satisfies Q}\}$,

S_1: $\{Z \mid Z \in S$ and two of the three points of Z in Γ_{u-}, with the remaining points in $\Gamma_{u+}\}$, and

S_2: $\{Z \mid Z \in S$ and two of the three points of Z in Γ_{u+}, with the remaining point in $\Gamma_{u-}\}$.

Finally define,

h(b;Z): height at b of the hyperplane determined by a triple Z.

We are ready to demonstrate,

Lemma 5.3.1: For any triple $Z \in S_1$, there is a triple $Z' \in S_2$ such that $h(b;Z') \geq h(b;Z)$.

Proof: Fix a triple $Z = \{z_1, z_2, z_3\} \in S_1$. Assume $z_1, z_2 \in \Gamma_{u-}$ with $x_{11} < x_{12}$. Then $h(b;Z) = z_{3u} \cdot \|y - b\| / (\|y - z_{3\ell}\|) + \theta \cdot \|b - z_{3\ell}\| / (\|y - z_{3\ell}\|)$[1], where y is the intersection of $\ell(b, z_3)$ with $\Gamma_{\ell-}$ and θ is the height of $\ell(z_1, z_2)$ at y. Now let $z_2' = (0,0,G(0))$ and consider a family $\{Z'\}$ of triples, where $Z' = \{z_1', z_2', z_3\}$ and z_1' is such that $x_{11}' \leq x_{11}$. Clearly $Z' \in S_2$. It is to be shown that we can choose z_1', depending on the magnitude of θ, so that $h(b;Z') \geq h(b;Z)$.

i) $\theta = G(0)$. Choose $z_1' = z_1$, and we have $h(b;Z') = h(b;Z)$.

ii) $\theta < G(0)$. By the analogous consideration as above, $h(b;Z') = z_{3u} \cdot \|y - b\| / (\|y - z_{3\ell}\|) + \theta' \cdot \|b - z_{3\ell}\| / (\|y - z_{3\ell}\|)$, where θ' is the height of $\ell(z_1', z_2')$ at y. But $\theta' = z_{1u} \cdot \|y\| / \|z_{1\ell}'\| + G(0) \cdot \|z_{1\ell}' - y\| / \|z_{1\ell}'\|$ and this can be made as near G(0) as we please by choosing z_1', such that $\|z_{1\ell}'\|$ is large enough since $G(t) \geq 0$. Finally, since θ' can be made greater than θ, the lemma is proved.

---

[1] Here and on, $\|x\|$ denotes the E_2-norm of x and $\ell(x,y)$ denotes the line through the points x and y.

It is immediate that $\sup_{Z\in S} h(b;Z) = \sup_{Z\in S_2} h(b;Z)$ from this lemma. The next step is to replace the set S_2 of triples by a set R of pairs of distinct points $\{z_1,z_2\}$, one each from Γ_{u+} and Γ_{u-}.

For this, let Y represent a pair $\{z_1,z_2\}$ of distinct points in Γ_u and state the condition Q' on Y;

Condition Q': $b\in CH[Y_\ell]$.

Define $R = \{Y \mid Y \text{ satisfies } Q'\}$.

Lemma 5.3.2: Under the assumption that $G(\cdot)\in\mathcal{G}$ defined in 5.2.3,

$$\sup_{Z\in S_2} h(b;Z) = \sup_{Y\in R} h(b;Y).$$

Proof: i) For any pair $Y\in R$, there is a triple $Z\in S_2$ such that

$$h(b;Y) = h(b;Z) \Longrightarrow \sup_{Z\in S_2} h(b;Z) \geq \sup_{Y\in R} h(b;Y).$$

ii) Fix a triple $Z\in S_2$. Assume $z_2,z_3\in\Gamma_{u+}$ with $x_{12} < x_{13}$. Let y be the intersection of $\ell(b,z_{1\ell})$ with $\Gamma_{\ell+}$, and θ be the height at y of $\ell(z_2,z_3)$. Also consider the pair $Y = \{z_1,z'\}$, where z' is on Γ_{u+} such that $z'_\ell = y$. Clearly $Y\in R$. We let $r_1 = \|b - y\| / \|z_{1\ell} - y\|$ and $r_2 = \|z_{1\ell} - b\| / \|z_{1\ell} - y\|$. Then,

$$h(b;Y) = z_{1u} \cdot r_1 + z'_u \cdot r_2 = z_{1u} \cdot r_1 + G(\gamma t_2 + (1-\gamma)t_3) \cdot r_2$$

$$\geq z_{1u} \cdot r_1 + [\gamma G(t_2) + (1-\gamma)G(t_3)] \cdot r_2$$

$$= z_{1u} \cdot r_1 + \theta \cdot r_2 = h(b;Z),$$

where the inequality is due to the fact that $G(\cdot)\in\mathcal{G}$ and substitutions for $\gamma = \|y - z_{3\ell}\| / \|z_{2\ell} - z_{3\ell}\|$, and for $G(t_2) = z_{2u}$, $G(t_3) = z_{3u}$. In fact we have shown that, for any triple $Z\in S_2$, there is a pair $Y\in R$ such that $h(b;Y) \geq h(b;Z)$. Hence the lemma follows.

We proceed to compute $h(b;Y)$ for any $Y\in R$. To a pair $Y = \{z_1,z_2\}$, we associate the acute angle α between $\ell(z_1, z_2)$ and $\Gamma_{\ell+}$, and note that $Y\in R \Longleftrightarrow \alpha\ [0,\pi/2]$. (We also use notations like $h(b;\alpha)$, $h(b;f(\alpha))$ for $h(b;Y)$ in view of this.)

For a point $b = (b_1,b_2)'$,

$$h(b;\alpha) = G(-d - s \tan\alpha) \cdot d/(d + s \tan\alpha) +$$

$$G(s + d \cot\alpha) \cdot s \tan\alpha/(d + s \tan\alpha),\ \alpha\in[0,\pi/2],$$

where $d = (b_2 - b_1)/2$, $s = (b_1 + b_2)/2$.

Reparametrizing by $p = d/(d + s \tan\alpha)$,

$$h(b;p) = p \cdot G(-d/p) + (1 - p) \cdot G(s/(1 - p)),\ p\in(0,1).$$

In view of the fact that $G(-d/p) \leq G(0) \leq G(s/(1 - p))$, $\forall\ p\in(0,1)$,

we conclude that $\sup_{p\in(0,1)} h(b;p) = G(s)$.

In summary, we may state that

1. The value of the program P_3 is given by $G((b_1 + b_2)/2)$.

Unless $b_1 = b_2$, in which case there does not exist an extremal (honest) distribution achieving this program value. If $b_1 = b_2$, i.e., $b\in\Gamma_{\ell+}$, the extremal distribution is degenerate at b_1.

2. Let $\mathcal{L}$ be the line parallel to $\Gamma_{\ell-}$. Then for every $b = (b_1,b_2)\in\mathcal{L}$, the program value is the same.

3. In case $G(t)$ is not concave in $t\in[0,\infty)$, define

$$G_c(t) = G(t),\ t\ (-\infty,0),$$

$$= \text{lowest concave function} \geq G(t),\ t\in[0,\infty).$$

Then, the program value is given by $G_c((b_1 + b_2)/2)$.

Ex. b. (Ex. 1 of section 2).

$$P_3: \sup_{x\in C\cap\mathcal{L}_b} x_3$$

where $C = CH(\Gamma)$,

$\mathcal{L}_b = \{x = (x_1,x_2,x_3) \mid x_i = b_i,\ i = 1,2,\ x_3\in E_1\}$ with $b = (b_1,b_2)$, and $G(\cdot)\in\mathcal{G}$ defined in 5.2.3.

As before, $\Gamma = \{(x_1,x_2,x_3) | x_1 = t,\ x_2 = t^2,\ x_3 \leq G(t);\ \text{some } t \epsilon E_1\}$

$$\Gamma_u = \{(x_1,x_2,x_3) | x_1 = t,\ x_2 = t^2,\ x_3 = G(t);\ \text{some } t \epsilon E_1\}$$

$$\Gamma_\ell = \{(x_1,x_2,0) | x_1 = t,\ x_2 = t^2,\ \text{some } t \epsilon E_1\}$$

As in the analysis of Ex. a, the reduction of the collection of triples in Γ_u to the collection of pairs in Γ_u can be made based on the fact that

i) pick at most one point in Γ_{u-},

ii) pick one point in Γ_{u+} instead of two, with the assumption of concave Γ_{u+}, which fact may be deducible from the concavity of the upper half c.d.f. and the convexity of the function t^2.

Using the analogous argument, we begin with a collection of pairs $Y = \{z_1,z_2\}$, where $z_1 \epsilon \Gamma_{u-}$, $z_2 \epsilon \Gamma_{u+}$, satisfying the condition Q', i.e., restrict attention to the set $R = \{Y | Y \text{ satisfies } Q'\}$. We reparametrize $Y \epsilon R$ by $\theta \epsilon [0,\pi]$, where the angle θ is measured between $\ell(z_{1\ell},z_{2\ell})$ and the x_1-axis. For given $b = (0,b_2) \epsilon CH[\Gamma_\ell]$, by letting $p = \frac{1}{2} \tan \theta$, we find

$$h(b;p) = G(p + \sqrt{p^2 + b_2}) \cdot \left\{(\sqrt{p^2 + b_2} - p)/2\sqrt{p^2 + b_2}\right\} +$$

$$G(p - \sqrt{p^2 + b_2}) \cdot \left\{(\sqrt{p^2 + b_2} + p)/2\sqrt{p^2 + b_2}\right\}, \text{for } p \neq \infty$$

$$= G(0), \text{ otherwise.}$$

One can check that $h(b;p)$ is convex on $(0,\infty)$ and concave on $(-\infty,0]$ in view of the fact that $G(\cdot)\in\mathcal{G}$. From the experience with Ex. a, it is enough to consider $p\in(-\infty,0]$ to compute $h^*(b) = \sup_{\theta} h(b;\theta) = \sup_{p} h(b;p)$. By differentiating $h(b;p)$ with respect to p and setting it equal to 0, we get,

$$g(p + \sqrt{p^2 + b_2}) + g(p - \sqrt{p^2 + b_2}) = (\sqrt{p^2 + b_2})^{-1}\{G(p + \sqrt{p^2 + b_2}) - G(p - \sqrt{p^2 + b_2})\}, \qquad (5.3.1)$$

where $g(\cdot)$ is the density function of $G(\cdot)$. The program value can be obtained once the equation 5.3.1 is solved for p. Note that 5.3.1 yields the equation (8) in section 2, when we let $t_2 = p + \sqrt{p^2 + b_2}$, $t_1 = p - \sqrt{p^2 + b_2}$ and $G(\cdot)$ be the Laplace distribution assumed there. It is interesting to note that the relation 5.3.1 has a certain geometric interpretation, i.e., p is to be determined such that the area under the density function $g(\cdot)$ between two points t_1 and t_2 is equal to the area of the right angle trapezoid formed by the four points t_1, $g(t_1)$, $g(t_2)$, and t_2.

4. Conclusion

Some additional remarks and comparison of the two methods of sections 2 and 3 are in order. Implicitly assumed in the use of the weak duality method of section 2 is that the program pair (P,D) is well-formulated in the sense that the solution pair (F^*,λ^*) is guaranteed to exist, and candidate solution pairs are easily accessible to provide the test for optimality. This assumption is satisfied in Ex. 1, but not in Ex. 2 except when $b_1 = b_2$. More precisely, in Ex. 2 with $b_1 \neq b_2$, there does not exist an extremal distribution F^* even though we have equality of values (in the "sup" and "min" sense) of the program pair (P,D), together with an optimal solution λ^* for D. This fact is attributable to lack of closure of $C \cap \Sigma_b$ (notation of P_3). Therefore, the weak duality method is not judged to be appropriate when situations like those of Ex. 2 prevail. In general, however, we may conclude that the weak duality method is computationally superior in view of its algorithmic nature, while the method of section 3 perhaps provides greater geometric insight. For example, the fact that there is no duality gap in both Ex. a and

Ex. b for any b$\in$Boundary of CH[h(T)], and consequently there is a solution λ^* for D, can be explained by the "corner" between the "upper boundary" and relevant "side" of the set C which makes it possible to single out at least one nonvertical hyperplane at the boundary.

VI. BIBLIOGRAPHY

1. Bahadur, R. R. 1971. Some limit theorems in statistics. Regional conference series in applied mathematics 4. S.I.A.M., Philadelphia, PA.

2. Bahadur, R. R. 1978. Private communication. Dept. of Statistics, University of Chicago, Chicago, IL.

3. Bahadur, R. R., and Raghavachari, M. 1971. Some asymptotic properties of likelihood ratios on general sample spaces. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability I:129-152.

4. Bahadur, R. R., and Rao, R. R. 1960. On deviation of the sample means. Ann. Math. Statist. 31:1015-1027.

5. Bahadur, R. R., and Zabell, S. L. 1978. Large deviations of the sample mean in general vector space. To appear in the Annals of Probability.

6. Balakrishnan, A. V. 1968. Basic concepts of information theory. Chapter 5 in A. V. Balakrishnan et al., eds. Communication theory. Inter-university Electronics Series, Vol. 6. McGraw-Hill, New York, NY.

7. Berger, A. 1951. Remarks on separable spaces of probability measures. Ann. Math. Statist. 22:119-120.

8. Berger, T. 1971. Rate distortion theory. Prentice-Hall, Inc., Englewood Cliffs, NJ.

9. Boza, L. B. 1971. Asymptotically optimal tests for finite Markov chains. Ann. Math. Statist. 42:1992-2007.

10. Chernoff, H. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Statist. 23:493-507.

11. Chernoff, H. 1972. Sequential analysis and optimal design. Regional Conference Series in Applied Mathematics 8. S.I.A.M., Philadelphia, PA.

12. Church, J. D., and Harris, B. 1970. The estimation of reliability from stress and strength relationship. Technometric 12:49-54.

13. Conn, P. W. 1969. Asymptotic properties of sequences of positive kernels. Unpublished Ph.D. Thesis. Iowa State University.

14. Cramér, H. 1938. Sur un nouveau théorème-limite de la théorie des probabilités. Act. Sci. et Ind. 736:5-23.

15. Csiszár, I. 1967. Information-type measures of difference of probability distributions. Studia Sci. Math. Hungar. 2:299-318.

16. David, H. T., and Kim, Geung-Ho. 1978. Pragmatic optimization of information functionals. To appear in the Proceedings of the International Conference of Optimization and Statistics at Bombay, India, 1977.

17. Duffin, R. J. 1970. Duality inequalities of mathematics and science. pp. 402-423 in J. Rosen, O. Mangasarian, and K. Ritter, eds. Nonlinear Programming. Academic Press, New York, NY.

18. Efron, B. 1978. The geometry of exponential families. Ann. Statist. 6:362-276.

19. Feller, W. 1971. An introduction to probability theory and its applications. Vol. II, 2nd edition. Wiley, New York, NY.

20. Govindarajulu, Z. 1968. Distribution-free confidence bounds for P(X<Y). Ann. Inst. Stat. Math. 20:229-238.

21. Harris, T. E. 1963. Theory of Branching Processes. Springer-Verlag, Berlin.

22. Hendrickson, A. D., and Buehler, R. J. 1971. Proper scores for probability forecasters. Ann. Math. Statist. 42:1916-1921.

23. Hoeffding, W. 1965. Asymptotically optimal tests for multinomial distributions. Ann. Math. Statist. 36:369-408.

24. Hoeffding, W. 1967. On the probabilities of large deviations. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1:203-219.

25. Isii, K. 1964. Inequalities of the types of Chebyshev and Cramér-Rao and mathematical programming. Ann. Inst. Stat. Math. 16: 277-293.

26. Karlin, S., and Studden, W. J. 1966. Tchebycheff systems: With applications in analysis and statistics. Interscience Publishers, New York, NY.

27. Kemperman, J. H. B. 1968. The general moment problems, a geometric approach. Ann. Math. Statist. 39:93-122.

28. Kerridge, D. F. 1961. Inaccuracy and inference. J. Roy. Stat. Soc., Ser. B, 23:184-194.

29. Khinchin, A. I. 1949. Mathematical foundations of statistical mechanics. Dover, New York, NY.

30. Kiefer, J., and Wolfowitz, J. 1959. Optimum designs in regression problems. Ann. Math. Statist. 30:271-294.

31. Kim, Geung-Ho, and David, H. T. 1978. Bivariate distributions as saddle points of mutual information. To appear in J. of Appl. Prob.

32. Kim, Geung-Ho, and David, H. T. 1978. Large deviations of functions of Markovian transitions and mathematical programming duality. Submitted for publication.

33. Kim, G. H., El-Sabbagh, M. F. A., and David, H. T. 1977. Lagrangian duality and large deviations for Markov chains (preliminary report). IMS Bulletin 6:140.

34. Kingman, J. F. C. 1963. On inequalities of the Tchebychev type. Proc. Cambridge Phil. Soc. 59:135-146.

35. Kolmogorov, A. N. 1956. On the Shannon theory of information transmission in the case of continuous signals. IRE Trans. Infor. Th. IT-2:102-108.

36. Koopmans, L. H. 1960. Asymptotic rate of discrimination for Markov processes. Ann. Math. Statist. 31:982-994.

37. Lanford, O. E., and Ruelle, D. 1969. Observables at infinity and states with short range correlations in statistical mechanics. Comm. Math. Phys. 13:194-215.

38. Lehmann, E. L. 1959. Testing Statistical Hypotheses. Wiley, New York, NY.

39. Lindley, D. V. 1956. On a measure of the information provided by an experiment. Ann. Math. Statist. 27:986-1005.

40. Madsen, R. W., and Conn, P. S. 1973. Ergodic behavior for non-negative kernels. Ann. Prob. 1:995-1013.

41. Meeks, D. H., and Francis, R. L. 1973. Duality relationships for a nonlinear version of the generalized Neyman-Pearson problem. J. Optim. Th. Appl. 11:360-378.

42. Mischke, C. R. 1976. Private communication. Dept. of Mechanical Engineering, Iowa State University, Ames.

43. Montroll, E. W. 1947. On the theory of Markov chains. Ann. Math. Statist. 18:18-36.

44. Mulholland, H. P., and Rogers, C. A. 1958. Representation theorems for distribution functions. Proc. London Math. Soc., Series 3, 8:177-223.

45. Noble, B., and Sewell, M. J. 1972. On dual extremum principles in applied mathematics. Transactions of the 17th Conference of Army Mathematicians 17:617-737.

46. Preston, C. 1976. Random fields. Lecture Notes in Math. Vol. 534. Springer-Verlag, New York, NY.

47. Pyne, D. A. 1972. Duality in abstract mathematical programming with applications to statistical problems. Unpublished Ph.D. Thesis. Iowa State University.

48. Richter, H. 1957. Parameterfreie Abschätzung und Realisierung von Erwartungswerten. Blätter der Deutschen Gesellschaft für Versicherungs-mathematik 3:147-161.

49. Ritter, K. 1967. Duality for nonlinear programming in a Banach space. SIAM J. Appl. Math. 15:294-302.

50. Robbins, H. 1948. Mixture of distributions. Ann. Math. Statist. 19:360-369.

51. Rockafellar, R. T. 1974. Conjugate duality and optimization. Regional conference series in applied mathematics 16. S.I.A.M., Philadelphia, PA.

52. Rogosinski, W. W. 1958. Moments of non-negative mass. Proc. Roy. Soc. (London), Ser. A, 245:1-27.

53. Royden, H. L. 1968. Real analysis. The Macmillan Company, New York, NY.

54. Sanov, I. N. 1957. On the probability of large deviations of random variables. (Russian) Mat. Sbornik N.S. 42:11-44. English translation: Select. Transl. Math. Statist. and Prob. 1(1961): 213-244.

55. Sethuraman, J. 1961. Some limit theorems for joint distributions. Sankyha 23A:379-386.

56. Spitzer, F. 1971. A variational characterization of finite Markov chains. Ann. Math. Statist. 43:303-307.

57. Sposito, V. A. 1970. Aspects of duality in linear programming. Unpublished Ph.D. Thesis. Iowa State University.

58. Sposito, V. A., and David, H. T. 1971. Saddle point optimality criteria of nonlinear programming problems over cones without differentiability. SIAM J. Appl. Math. 20:698-702.

59. Van Slyke, R. M., and Wets, R. J.-B. 1968. A duality theorem for abstract mathematical programs with applications to optimal control theory. J. Math. Anal. Appl. 22:679-706.

60. Varaiya, P. P. 1967. Nonlinear programming in Banach space. SIAM J. Appl. Math. 15:284-293.

61. Wets, R. J.-B. 1970. Necessary and sufficient conditions for optimality: A geometric approach. Operations Research-Verfahren 8:305-311.

62. Whittle, P. 1971. Optimization under constraints; theory and applications of nonlinear programming. Wiley-Interscience, New York, NY.

VII. ACKNOWLEDGMENTS

The essential part of this research is supported by a grant from the Air Force Office of Scientific Research. Professor H. T. David directed the research, and made substantial contributions to the results recorded here. His characteristically inspiring guidance extended over the various stages of my graduate study has been vital to my professional growth.

For opportunities of getting invaluable training in the field of industrial engineering, I am indebted to Professor K. L. McRoberts. On several occasions, he has been a provider of stimulating problems, together with the necessary support to carry out the relevant investigations.

All of the staff of the Statistical Numerical Analysis and Data Processing Section of the Statistics Lab have been extremely helpful. In particular, I am very grateful to Professor V. A. Sposito for his continuous encouragement as well as his generously sharing of his expertise on numerous occasions. Also, I am obliged to Professor W. J. Kennedy, who helped me start out in computing, for his continuous flow of expert advice during my subsequent years in the Section.

To the other members of my committee, Professor R. A. Groeneveld, Professor D. L. Isaacson, and Professor H. D. Meeks, I also would like to express my thanks for their generously giving of their time and helpful comments.

As for typing of the dissertation, Jean Bodensteiner performed a marvelous job, which I appreciate very much.